Apprentissage de vote de majorité pour la classification supervisée et l'adaptation de domaine : approches PAC-Bayésiennes et combinaison de similarités

Emilie Morvant

Aix*Marseille Université

École doctorale mathématique et informatique - ED 184

Laboratoire d'Informatique Fondamentale

UMR 7279 - CNRS

Thèse

présentée pour obtenir le grade de

Docteur d'Aix*Marseille Université

Spécialité : Informatique

par

Emilie Morvant

Apprentissage de vote de majorité pour la classification supervisée et l'adaptation de domaine :

approches PAC-Bayésiennes et combinaison de similarités

Thèse soutenue publiquement le 18 septembre 2013

Jury :

M^me^ Michèle Sebag	Directrice de recherche CNRS, Université d'Orsay	*Rapportrice*
M. Mario Marchand	Professeur, Université Laval, Québec	*Rapporteur*
M. Antoine Cornuéjols	Professeur, AgroParisTech	*Examinateur, Président*
M. Rémi Gilleron	Professeur, Université de Lille	*Examinateur*
M. Liva Ralaivola	Professeur, Aix*Marseille Université	*Examinateur*
M. Amaury Habrard	Professeur, Université de Saint-Étienne	*Directeur*
M. Stéphane Ayache	Maître de Conférences, Aix*Marseille Université	*Co-Directeur*

À Jade

Remerciements

Bon, il paraît qu'il faut remercier la terre entière une fois sa thèse finie... Alors c'est parti et je dirais même plus : merci à tout l'univers de m'avoir amener jusqu'à ce jour ! Et tout particulièrement...

Tout d'abord, merci Amaury — en tant que "presque" vrai Stéphanois — tu m'as accueillie à bras ouverts dès mon arrivée à Marseille ! Merci de ne pas m'avoir laissée seule au milieu de tous ces Marseillais. Tu as su me guider à merveille dans le monde de la recherche et, sans toi, il est certain que je n'en serais pas là. J'espère que les liens que nous avons tissé dureront dans le temps et dans la recherche. Merci à toi aussi Stéphane de m'avoir laissé suivre ma voie loin du multimédia. Je t'en suis vraiment reconnaissante. Maintenant que mes chefs y sont passés (ça... c'est fait), j'aimerais remercier les membres de mon jury d'avoir accepté de juger ma thèse durant l'été. Je tiens à remercier Mario Marchand, professeur à l'université Laval de Québec, et le fait qu'il ait survécu à tous les classifieurs (plutôt de classificateurs) présents dans cette thèse ! Mais tabarnak ! Trêve de plaisanteries, je voudrais en particulier te remercier pour l'ensemble des échanges que l'on a pu avoir, que ce soit en conférence, au GRAAL à Québec, ou bien sur pour tes retours "d'examinateur externe de la thèse". J'aimerais également remercier Michèle Sebag, directrice de recherche CNRS à l'université d'Orsay, d'avoir elle aussi accepté de rapporter ma thèse. Merci pour vos retours des plus constructifs qui ont permis d'améliorer la qualité de cette thèse. Je remercie également Rémi Gilleron, professeur à l'université de Lille, et Antoine Cornuéjols, professeur à AgroParisTech, d'avoir eu la gentillesse d'accepter d'être examinateurs. Je vous suis, à tous, reconnaissante d'avoir consacré du temps à l'évaluation de mes travaux.
Mais non... Liva, je ne t'ai pas oublié (je gardais le meilleur pour la fin hein ? !). Bon, ok, tu n'as jamais été mon directeur, mais je te dois beaucoup et je te considère comme mon grand frère scientifique ! Je te remercie pour ta présence, ton écoute, tes conseils pertinents et ta sagesse de "vieux". Je n'oublierais jamais le défi — Ô combien réussi — d'écrire un papier en une semaine pour ICML ! Et quel papier ? ! c'est quand même grâce à lui que je me suis réellement laissée prendre au jeu du PAC-Bayes. C'était trop bon ! Enfin bref, je pourrais parler de pleins de choses de la vie, mais je vais simplement rajouter merci d'avoir bien voulu être examinateur de ma thèse et d'avoir lu ses 250 pages (enfin 260 maintenant) !

Après le jury, j'aimerais remercier tout Qarma and co. *a.k.a.* The couloir ! Sans vous tous, ces trois années auraient été bien fades... Cécile, la présence féminine de l'équipe : ça aurait vraiment été difficile sans toi, entourée de toute cette testostérone. François avec qui je n'ai jamais eu l'occasion de faire de la recherche : peut-être qu'un jour je me mettrais aux méhodes spectrales... Valentin qui a partagé un stagiaire pendant l'été ! C'était cool. F.-X., dommage que tu n'es pas eu le temps de m'initer à l'astrophysique. Hachem qui a réussi à me faire "aimer" les noyaux à valeurs opérateurs alors que mon cœur est pris par l'(ϵ,γ,τ)-gooditude... Rémi, l'expatrié, mais qui n'en n'oublie pas son âme Stéphanoise : je me souviens d'une magnifique photo "Allez les verts" pour la finale de l'année. J'en ai fini avec le Qœur et je passe maintenant à la sous-

branche : les matheux ! Merci à vous tous : Caroline, Clotilde, Claire, Bruno, Marie-Christine, Frédéric, Thomas, Benjamin, Sébastiano, Sangnam, Alexandre, Pierre, Alain. Un merci spécial à Sandrine, merci à toi de m'avoir voiturée un bon nombre de fois ! Les permanents y étant tous passés (du moins, je l'espère...) je peux passer à l'âme de ce couloir : les (post-)doctorants ! Thomas et son magnifique enregistrement dans un taxi barcelonais mais — Dieu merci – perdu au fin fond de l'univers ! Je n'oublie pas le Petit Nice que l'on doit partager lorsque je pourrais t'appeler docteur à ton tour. Pierre et sa mémorable fin du monde : on ne pouvait pas rêver mieux pour une telle journée ! Au passage, je réponds à tes propres remerciements pour éviter toute "confusion" : si je t'ai abreuvé de cachets "anti-gueule-de-bois"... c'était bien parce que tu en avais besoin... Juliette sans qui mon mac n'aurait pas eu de gommettes ! Merci aussi d'avoir risqué ta vie un grand nombre de matin dans les quartiers Nord ! Emilie (l'autre) dont la vie de jeune fille a été enterrée avec brio le lendemain de la soumission de mon manuscrit. Nous partageons, en plus de notre prénom, un enregistrement plutôt intéressant de Thomas... une traversée mythique d'un pont de singe, ainsi qu'une future galette des rois en janvier. Mattias, un des derniers arrivés, mais pas des moindres... un Stéphanois de l'UJM ! Dire qu'il aura fallu attendre d'être à Marseille pour se parler pour la première fois. Un grand merci à vous 5 car ces 3 années n'auraient pas été les mêmes sans vous et surtout pour l'initiation à la grimpe (enfin pas pour Pierre) ! Sokol, nous avons partagé un bureau en tête à tête mais aussi la confusion de Liva et Pierre. Harold, que j'ai aimé tes moments de réflexions intenses, mais aussi fin du monde au Polikarpov ! Guillaume S. on n'a jamais été en thèse ensemble, mais merci quand même pour l'initiation à la Mauresque au Petit Nice ! Raphaël, dire qu'au début j'avais peur de toi ! Il est bien loin se temps-là ;-) Sylvain qui m'a refusé la chance d'avoir mon cerveau en 3D, mais c'est pas grave, je t'en veux pas. Antoine et Ugo et notre périple interminable de deux jours vers South Lake Tahoe (merci Antoine pour ton déo). Anaïk, merci pour ta bonne humeur naturelle ! Guillaume R., Julien et Hongliang, les derniers arrivés, nous n'avons pas trop pu passer trop de temps ensemble mais j'espère que nous aurons de nombreuses occasions pour combler ce manque en conférence ou ailleurs ! Je remercie également l'ensemble du LIF et ses "Angels" : Martine, Sylvie et Nadine. Je voudrais finir avec tous les p'tits jeunes avec qui j'ai fait des courses mémorables dans le couloir : Noé, Margot, Soléa, et Rachel !

Bon, ce n'est un secret pour personne, je ne peux pas m'arrêter là, il me faut parler de la branche Stéphanoise ! Je suis fière d'avoir entretenue la flamme qu'il existe entre Marseille et Sainté... Je me souviens de plusieurs conf' mémorables avec les Stéphanois et — en particulier — d'un meeting Lampada à Lille avec un fou rire interminable devant l'hotel... En premier lieu, je tiens à remercier François Jacquenet : sans lui, je n'aurais jamais candidaté en M2R à Marseille. Je voudrais remercier Marc S. de m'avoir laissée partir à Marseille "lorsque j'en ai eu besoin" (2 semaines avant la rentrée scolaire) ! Mais aussi de m'avoir laissée revenir "lorsque j'en ai eu besoin - le retour" (6 mois avant la fin de thèse). J'aimerais aussi te remercier pour tes conseils en LaTex lorsque je n'étais qu'une novice et, bien entendu, pour tous les échanges scientifiques (et moins scientifiques) que nous avons eu, en espérant qu'ils perdurent encore longtemps ! Bien sur je

remercie toute la team stéphanoise avec une pensée spéciale pour Elisa (et son estomac sur pattes), Baptiste (et sa bonne humeur permanente), Léo (et son accent espagnol), Michaël (et la lecture de ma thèse qu'il a du faire pendant les vacances), Mathias (et sa folie pour la course) , Marc B. (et les échanges sur facebook que l'on a pu avoir) ! Sans oublier J.-P. et Aurélien avec qui j'ai — entre autre — partagé l'(ϵ,γ,τ)-gooditude et Amaury (si on rajoute Mattias, je me demande comment il a fait pour nous supporter d'ailleurs) ! Merci Aurélien de m'avoir "appris" à faire des scripts et pour le magnifique ouvrage que tu m'as légué à ton départ. J.-P., j'espère que tu me pardonneras de t'avoir interdit à plusieurs reprise d'ouvrir la fenêtre… Au fait, le papier Aurélien, Emilie, J.-P. on l'attend toujous non ?

Enfin, un grand merci à la team Québécoise de m'avoir permis de trouver le GRAAL (qui est en permanence sur mon bureau) ! Un merci spécial à Pascal et J.-F. pour nos collaborations ! Et bien sûr, merci à François L. et à toute sa famille pour l'accueil super chaleureux que vous m'avez réservé durant mon séjour !

D'un point de vue plus personnel, je voudrais aussi remercier tous les non-chercheurs de m'avoir soutenue et supportée pendant toutes ces années. Luce, Annette, Deydey et Dam's, les Saint-Louisards ! Au fait, Luce, Annette je suis docteur avant vous : nananère ! Et Dam's, je n'oublierais pas le concert de Madonna à Québec, ma thèse n'aurait pas été la même sans cette expérience inoubliable ! ! Célyn, Flo, Val, Greg, Yo, Davide, Mimi, Tata (même si elle est loin) qui ont partagé les joies de l'UJM, des maths et de l'info avec moi, à l'époque où nous nous surnomions La famille Bouar avec Vachette, Biquette, La Chienne, Bouriquet, Poussin, Annyo, j'en passe et des meilleurs ;-). Un spécial merci à Davide et Fanny, d'une part vous m'avez offert le premier mariage parmi mes amis, et d'autre part pour le superbe hotel que vous me proposez lorsque je viens à "Gre" (enfin maintenant à Renage) ! Merci au Grup *a.k.a.* Nounours, Manu, Papy, Romain, Nico, Lucas, Eric, Amandine, Aude, Marlène, Alex, Aurore, Audrey, Amanda, Tounesse, Camille. Merci aussi à Elo' et Martine de m'avoir toujours soutenue et poussée à faire ce que j'aime. Un spécial "Kung-Fu" thanks à l'ami Djé (Gégé pour les intimes) et toute ta petite famille ! Toi qui a basculé du côté obscur du privée ! Sans toi, ces dernières années n'auraient pas été les mêmes (et même ces derniers mois si tu vois ce que je veux dire…) ! Et t'as vu ? je me suis même prise au jeu du Kung-Fu (ouaaaaaayyyyyyaaaaaaa ← ça c'est ce que j'appelle un Kiai de fou).

Un grand merci à toute ma famille ! Ma mère, Michmich, qui commence à comprendre ce que je fais… (au fait Liva, elle t'as trouvée mignon lol). Mes soeurs, Angie & Lulu (et à mes beaux-frères Fab — l'officiel — et Yohan — le moins-officiel—), vous aussi n'y comprennez surement rien mais je sais que m'encouragez comme des folles. Mon petit frère, Tom (alias Tominou pour les intimes, c'est plus swaggy), le BG-G33K de la famille (car oui, il y a pire que moi dans cette famille) ! Spéciale dédicace à Jade, mon plus jeune public, née le jour de la soumission de mon manuscrit (trop la classe quand même). Merci Cécile d'être entrée dans ma vie ces derniers mois, ma fin de thèse a été remplie de bonheur de joie et … Mes dernières pensées vont pour mon père, J.-L. M., qui est celui qui m'a transmis la chose la plus importante dans ce métier : la passion.

« *L'important n'est pas de convaincre*
mais de donner à réfléchir. »
Bernard Werber

Table des matières

Liste des figures

Liste des tableaux

Introduction

L'apprentissage automatique [1] est une discipline informatique considérée généralement comme l'un des champs d'étude de l'intelligence artificielle et se situe à la frontière de l'informatique et des mathématiques appliquées (statistiques et optimisation). La notion d'apprentissage englobe toute méthode permettant de construire un modèle le plus proche possible de la réalité à partir de données observées. Son objectif est d'extraire et d'exploiter automatiquement l'information présente dans des jeux de données, en s'intéressant au développement, à l'analyse et à l'implémentation de méthodes capables de s'améliorer à l'aide des observations. De ce fait, il se décline sous "un vaste champ de formes". Nous pouvons, entre autres, citer la classification dont une illustration est la différentiation automatique entre spams et hams (c'est-à-dire les e-mails désirés), ou la régression dont un exemple est la prédiction de la température du jour en fonction des informations relevées par les stations météo. Selon les données et les objectifs, de nombreuses applications existent, notamment en multimédia, en biologie, en traitement du signal, en traitement automatique de la langue, etc.

Dans cette thèse, nous nous plaçons dans le cadre de l'apprentissage statistique introduit par Vapnik et Chervonenkis, lié à la théorie d'études statistiques sur les processus empiriques. En particulier, nous nous intéressons au paradigme de la classification supervisée qui s'oppose à celui de la classification non supervisée (par exemple le *clustering*) pour laquelle nous ne disposons d'aucune supervision sur la classe des données observées. Reprenons l'exemple classique et concret d'un système de filtrage ham/spam. Pour ce faire, nous avons à notre disposition des observations : ce sont des messages déjà étiquetés, prenant la forme de couples $(\mathbf{x}, y)$, où $\mathbf{x}$ est la représentation d'un message, comme un vecteur de fréquences de mots dans le texte, et où y est la classe/l'étiquette affectée au message $\mathbf{x}$ et classiquement modélisée par -1 ou $+1$ (-1"=" spam, $+1$"=" ham). Le but est alors de trouver — d'apprendre — une fonction qui attribue automatiquement une de ces deux classes à un nouveau message, en commettant le moins d'erreurs possible. On parle alors de classification puisque la tâche est clairement de prédire la classe d'appartenance des données, et elle est dîte supervisée puisque l'apprentissage se base sur des observations déjà étiquetées. La classification supervisée se formalise alors comme l'apprentissage ou la construction d'une fonction, souvent appelée hypothèse ou classifieur, à partir d'un échantillon d'apprentissage composé d'observations indépendantes et identiquement distribuées

1. *Machine learning* en anglais. Nous invitons le lecteur à se référencer aux ouvrages suivants : [Mitchell, 1997, Bishop *et al.*, 2006, Cornuéjols et Miclet, 2010, Mohri *et al.*, 2012].

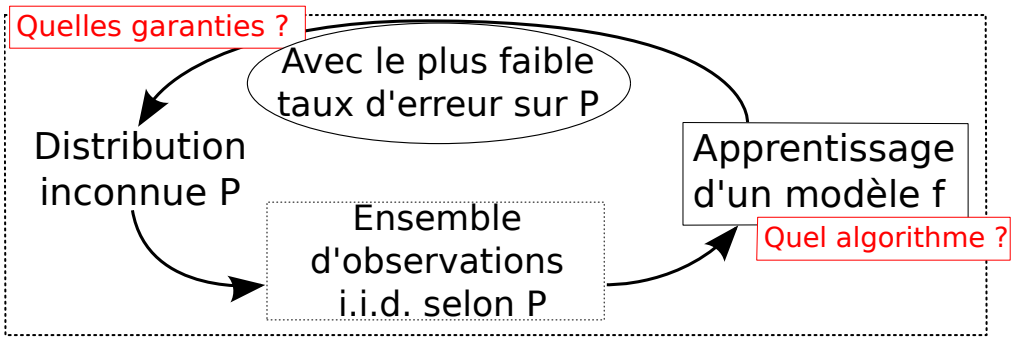


FIGURE Intro.1 – **Intuition des problématiques liées à l'apprentissage statistique** *: étant donnée une tâche, par quelle méthode et avec quelles garanties peut-on apprendre un modèle performant ?*

selon une distribution de probabilité fixée et inconnue. Un aspect important doit être souligné ici : le modèle appris doit se montrer performant sur les observations mais aussi — et surtout — sur les nouvelles données. Autrement dit, puisque la distribution des données est supposée inconnue, la question majeure en apprentissage statistique concerne l'apprentissage d'une hypothèse d'un classifieur se comportant le mieux possible sur l'ensemble de la distribution : il doit donc avoir de bonnes propriétés en généralisation (l'intuition est illustrée sur la figure Intro.1).

Dans la littérature, de nombreux algorithmes ont été proposés pour résoudre ces tâches de classification. La plus intuitive est la méthode des k plus proches voisins [Cover et Hart, 1967] qui consiste simplement à choisir la classe majoritaire parmi les k observations les plus proches de la donnée à étiqueter.
Nous pouvons également citer les méthodes faisant appel à des arbres de décision telles que [Breiman *et al.*, 1984, Quinlan, 1993], encore très populaires puisque facilement interprétables. Le classifieur prend, ici, la forme d'un arbre où une feuille correspond à un classe. Cette classe est obtenue en fonction des décisions prises à chaque étape du parcours de la branche.
Un autre type de méthode est celui des réseaux de neurones, dont les fondements ont été introduits par [Lettvin *et al.*, 1959]. Un réseau de neurones artificiel s'inspire des réseaux de neurones biologiques. Il est, en général, constitué d'une succession de couches interconnectées de neurones dits "formels" dont les entrées correspondent aux sorties de la couche précédente : chaque neurone est relié aux neurones de la couche précédente par des synapses calculant, classiquement, une somme pondérée des sorties de la couche précédente. Un des algorithmes historiques les plus simples, rentrant dans cette catégorie, est celui du perceptron [Rosenblatt, 1958]. Les extensions récentes de ces approches forment une thématique active et importante en apprentissage automatique appelée le *deep learning* [Bengio, 2009].
D'autres approches se focalisent sur l'apprentissage de combinaisons de classifieurs. On peut en particulier mentionner les algorithmes de type *boosting*, très populaires dans les années 90, dont l'un des plus utilisés est Adaboost [Freund et Schapire, 1996]. L'objectif du *boosting* est de combiner des classifieurs simples et faibles (c'est-à-dire faisant légèrement mieux que l'aléatoire). À chaque étape un classifieur est ajouté et les poids des observations sont modifiés afin de donner une plus grande importance

aux exemples mal classés.
Enfin, un des algorithmes les plus renommés est celui des machines à vecteurs de support [Boser *et al.*, 1992] dont le but est d'apprendre un classifieur linéaire étiquetant correctement les observations, tout en maximisant la distance entre le classifieur et ces observations. Ce concept, communément appelé la marge, permet de définir une notion de confiance sur les données. L'un des avantages de cette approche est qu'elle permet l'apprentissage de classifieurs non linéaires en reformulant le problème comme une classification linéaire dans un espace de projection défini à partir de fonctions noyaux. Nous détaillerons ce principe dans le premier chapitre de cette thèse.

Motivations Dans le contexte particulier dans lequel cette thèse s'est déroulée, la problématique étudiée est l'étiquetage automatique de vidéos, dans une situation d'indexation sémantique, dans le but de suggérer de nouvelles vidéos à l'utilisateur. De telles données multimédia sont habituellement représentées sous plusieurs formes. Basiquement, lorsque l'on parle de vidéos, nous pouvons à la fois décrire le son et l'image (et ce de différentes manières). Les données sont donc définies par un couple $(\mathbf{x}, y)$, où y est la classe d'appartenance du document vidéo (par exemple, l'absence ou la présence d'un objet particulier) et où cette fois-ci $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ est décomposé en n représentations distinctes de plusieurs informations du document. Autrement dit, la tâche d'apprentissage peut être vue comme un apprentissage de n classifieurs associés aux n couples $(\mathbf{x}_i, y)$, que l'on va devoir combiner. De plus, la personnalisation des suggestions soulève une seconde question importante. En effet, à l'instar d'un système de filtrage de spams, des vidéos intéressantes pour un utilisateur donné ne le seront pas nécessairement pour un nouvel utilisateur. Dans le même esprit, l'expansion d'Internet induit une très grande quantité de sources de données et donc de corpora différents. Il faut donc être capable d'adapter le modèle d'un utilisateur à un autre ou d'un corpus à un autre, comme suggéré par la figure Intro.2. Deux grandes questions nous intéressent donc dans ce mémoire :

- La combinaison de classifieurs : Comment peut-on apprendre une combinaison de classifieurs issus de différentes connaissances *a priori* prenant la forme, par exemple, de différentes sources d'informations ou de représentations, et ce avec de bonnes garanties en généralisation ?

- L'adaptation de domaine : Comment peut-on apprendre à partir de données sources — pour lesquelles on dispose d'observations étiquetées — un classifieur performant sur des données cibles — pour lesquelles peu (ou pas) d'étiquettes sont disponibles ? D'un point de vue statistique, le cadre de l'adaptation de domaine suppose que la distribution des observations sources diffère de celle des nouvelles données cibles.

Notre objectif principal est donc d'apprendre une combinaison performante et robuste tout en tirant bénéfice des différents classifieurs avec ou sans adaptation. Dans un premier temps, en se plaçant dans un cadre non adaptatif, nous nous interrogeons

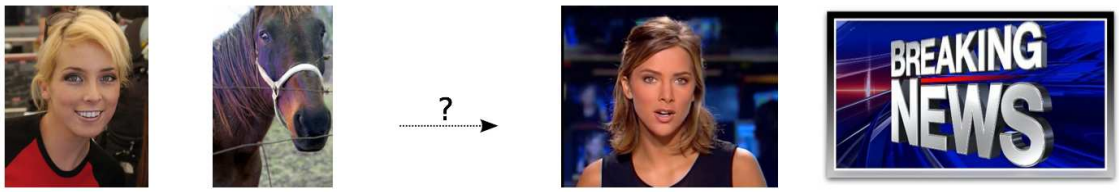


FIGURE Intro.2 – **Exemple d'une tâche d'adaptation** *: Supposons que la tâche de classification vise à identifier la présence d'un visage sur une image. À gauche, ce sont les observations étiquetées provenant d'un corpus d'images de type photographie. À droite, ce sont les nouvelles images à classer qui sont extraites de vidéos provenant d'un autre corpus. La question qui se pose est : comment adapter nos connaissances du corpus de gauche au corpus de droite ?*
Des travaux récents en traitement d'images ont montré que des classifieurs appris à partir de corpus différentes sont assujettis à des dégradations de performances [Torralba et Efros, 2011].

sur l'existence d'un cadre pertinent pour combiner des classifieurs. C'est ainsi que nous nous tournons logiquement vers la théorie PAC-Bayésienne [McAllester, 1999]. En effet, cette théorie offre un cadre naturel et élégant pour l'apprentissage d'une combinaison de classifieurs prenant la forme d'un vote de majorité pondéré en considérant une connaissance *a priori*. Nous l'exploitons donc pour proposer des avancées, plutôt théoriques, sur l'apprentissage de votes de majorité. Dans un second temps, nous répondons à la tâche de l'adaptation de domaine en s'inspirant des travaux fondateurs qui suggèrent de rapprocher les deux distributions dans un nouvel espace tout en gardant de bonnes garanties sur les observations sources [Ben-David *et al.*, 2007, Mansour *et al.*, 2009a, Ben-David *et al.*, 2010]. En ce sens, nous proposons d'adapter un espace de représentation défini à partir de similarités entre les données, permettant d'apprendre un vote de majorité sur ces similarités. Finalement, pour répondre aux deux problématiques simultanément, nous étudions l'adaptation de domaine avec un point de vue PAC-Bayésien dans le but de construire une combinaison de classifieurs capable de s'adapter à de nouvelles données.

Contexte de cette thèse. Cette thèse a été réalisée au sein de l'éQuipe d'AppRentissage automatique et MultimédiA (Qarma) du Laboratoire d'Informatique Fondamentale UMR CNRS 7279 d'Aix-Marseille Université entre octobre 2010 et septembre 2013. Les contributions présentées dans ce mémoire ont été développées dans le contexte du projet ANR VideoSense (ANR-09-CORD-026, `http://www.videosense.org/`) dont l'objectif est l'étiquetage automatique de vidéos par concepts de haut niveaux, d'évènements et émotions, en ciblant la recommandation vidéo et la monétisation de publicités, et de PASCAL2 (`http://pascallin2.ecs.soton.ac.uk/`), un réseau d'excellence européen portant sur l'apprentissage automatique, les statistiques et l'optimisation.

Organisation du mémoire. Cette thèse est organisée comme suit. La première partie présente les outils nécessaires à sa bonne compréhension :

- Le chapitre 1 introduit le contexte de base : la classification supervisée, la

dérivation de garanties en généralisation et différents algorithmes qui font références pour nos travaux.

- Le chapitre 2 présente un état de l'art non exhaustif de la problématique sur l'adaptation de domaine en apprentissage automatique, lorsque les distributions des observations et des nouvelles données diffèrent.
- Le chapitre 3 énonce la théorie PAC-Bayésienne se focalisant sur les votes de majorité pondérés sur un ensemble de classifeurs ou de fonctions.

La partie II présente nos contributions en théorie PAC-Bayésienne pour la classification supervisée sans adaptation :

- Dans le chapitre 4, nous nous focalisons sur l'apprentissage de votes de majorité pondérés sur un ensemble de classifieurs binaires dépendants des observations en tirant avantage d'une connaissance *a priori*. En ce sens, nous étendons l'algorithme MinCq [Laviolette *et al.*, 2011a] issu de la théorie PAC-Bayésienne. D'une part, nous en proposons une nouvelle formulation permettant de considérer une information *a priori* lors du processus d'apprentissage. D'autre part, nous en démontrons les garanties en généralisation dans le cadre des schémas de compression lorsque les classifieurs sont définis en fonction des observations. Après avoir énoncé notre extension, nous montrons empiriquement son intérêt pour la tâche de fusion de classifieurs en indexation sémantique de documents multimédia, ainsi que pour la combinaison de classifieurs de type plus proches voisins.
- Dans le chapitre 5, essentiellement théorique, nous énonçons la première borne en généralisation PAC-Bayésienne dans le cadre multiclasse qui se base sur la matrice de confusion. Concrètement, la mesure de risque que l'on considère est définie par la norme de la moyenne des matrices de confusion des classifieurs à combiner. Cette mesure, plus riche que le taux d'erreur usuel, offre un cadre théorique original pour s'attaquer à des tâches nécessitant une mesure plus informative qu'une mesure scalaire. Nous proposons, en outre, une analyse des relations qui unissent le moyennage sur classifieurs et le vote de majorité associé.

Nos contributions dans le cadre de l'adaptation de domaine pour la classification binaire sont présentées dans la partie III :

- Dans le chapitre 6, nous suivons l'intuition portée par les travaux fondateurs de la théorie de l'adaptation de domaine pour proposer une nouvelle méthode d'adaptation. Son principe est le suivant : nous cherchons à rapprocher les distributions dans un espace explicite défini par des fonctions de similarités dites (ϵ,γ,τ) - bonnes [Balcan *et al.*, 2008a, Balcan *et al.*, 2008b], tout en gardant de bonnes performances sur la distribution des observations. Dans un premier temps, nous ne faisons appel à aucune supervision sur les nouvelles données, puis nous étendons l'approche en supposant connues quelques étiquettes. Les

expérimentations menées démontrent la pertinence de l'approche pour l'adaptation de domaine.

- Dans le chapitre 7, nous dérivons la première analyse PAC-Bayésienne du problème de l'adaptation de domaine dans le cadre de la classification binaire sans information supervisée sur les nouvelles données. Notre contribution au cadre de l'adaptation de domaine repose sur la définition d'une nouvelle mesure de divergence entre distributions pertinente pour la théorie PAC-Bayésienne et donc pour les votes de majorités. Cette mesure nous permet de dériver une première borne PAC-Bayésienne d'adaptation de domaine qui a l'avantage d'être directement optimisable pour tout ensemble de classifieurs à combiner et nous en donnons une illustration pratique dans le cadre de classifieurs linéaires.

Enfin le lecteur pourra trouver dans les annexes la majeure partie des preuves des différents résultats obtenus, ainsi que quelques outils utiles à leur bonne compréhension. Notons que ces travaux ont donné lieu aux publications suivantes ainsi qu'à la participation à la tâche d'indexation sémantique dans le contexte des campages d'évalutations TrecVid 2011 et 2012. La participation 2011 est disponible en annexe.

Liste des publications

Journaux

Aurélien Bellet, Amaury Habrard, Emilie Morvant et Marc Sebban. Quadratic program for a priori constrained weighted majority vote - Application to nearest neighbor classifiers, 2013. **soumis**.

Emilie Morvant, Amaury Habrard et Stéphane Ayache. Parsimonious Unsupervised and Semi-Supervised Domain Adaptation with Good Similarity Functions. *Knowledge and Information Systems (KAIS)*, 33(2) :309–349, 2012. **publié**.

Conférences internationales

Pascal Germain, Amaury Habrard, François Laviolette et Emilie Morvant. PAC-Bayesian domain adaptation bound with specialization to linear classifiers. *International Conference on Machine Learning (ICML)*, 2013. **publié**.

Emilie Morvant, Sokol Koço et Liva Ralaivola. PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. *International Conference on Machine Learning (ICML)*, pages 815–822, 2012. **publié**.

Emilie Morvant, Amaury Habrard et Stéphane Ayache. Sparse domain adaptation in projection spaces based on good similarity functions. *IEEE International Conference on Data Mining series (ICDM)*, pages 457–466, 2011. **publié**, Selected as one of the best papers for possible publication in Knowledge and Information Systems (KAIS).

Workshops internationaux

Emilie Morvant, Stéphane Ayache, Amaury Habrard, Miriam Redi, Tanase Claudiu, Bernard Merialdo, Bahjat Safadi, Franck Thollard, Nadia Derbas et Georges Quenot. VideoSense at TRECVID 2011 : Semantic Indexing from Light Similarity Functions-based Domain Adaptation with Stacking. *TRECVID 2011 workshop*, 2011. **publié**.

Emilie Morvant, Amaury Habrard et Stéphane Ayache. On the usefulness of similarity based projection spaces for transfer learning. *Similarity-Based Patterns Recognition workshop (SIMBAD)*, pages 1–16, 2011. **publié**.

Conférences nationales

Aurélien Bellet, Amaury Habrard, Emilie Morvant et Marc Sebban. Vote de majorité a priori contraint pour la classification binaire : spécification au cas des plus proches voisins. *Conférence Francophone sur l'Apprentissage Automatique (CAp)*, 2013. **publié**.

Pascal Germain, Amaury Habrard, François Laviolette et Emilie Morvant. Une analyse pac-bayésienne de l'adaptation de domaine et sa spécialisation aux classifieurs linéaires. *Conférence Francophone sur l'Apprentissage Automatique (CAp)*, 2013. **publié**.

Emilie Morvant, Amaury Habrard et Stéphane Ayache. Étude de la généralisation de DASF à l'adaptation de domaine semi-supervisée. *Conférence Francophone sur l'Apprentissage Automatique (CAp)*, pages 111–126, 2012. **publié**.

Emilie Morvant, Stéphane Ayache et Amaury Habrard. Adaptation de domaine parcimonieuse par pondération de bonnes fonctions de similarité. *Conférence Francophone d'Apprentissage (CAp)* 2011. **publié**.

Communications

Pascal Germain, Amaury Habrard, François Laviolette et Emilie Morvant. PAC-Bayesian learning and domain adaptation. *Multi-Trade-offs in Machine Learning Workshop at NIPS*, Spotlight/Poster Presentation, 2012.

Emilie Morvant, Jean-Francis Roy, François Laviolette et Liva Ralaivola. Generalization of the C-bound to multiclass setting. *Women in Machine Learning Workshop (WiML)*, Poster, 2012

Emilie Morvant, Amaury Habrard et Stéphane Ayache. Sparse domain adaptation in a good similarity-based projection space. *Domain Adaptation Workshop at NIPS*, Poster Presentation, 2011.

Rapport de recherche

Emilie Morvant, Amaury Habrard et Stéphane Ayache. PAC-Bayesian majority vote for late classifier fusion. *ArXiv e-prints*, 2012. `http://adsabs.harvard.edu/abs/2012arXiv1207.1019M`.

Liste des notations

$X \in \mathbb{R}^d$	Espace d'entrée de dimension d
Y	Espace de sortie
Q	Le nombre de classes
P	Un domaine : distribution fixe et inconnue sur $X \times Y$
D	Distribution marginale de P sur X
$(\cdot)^\top$	Transposée d'un vecteur ou d'une matrice
$\mathbf{x} = (x_1, \ldots, x_d)^\top$	Vecteur de réels de dimension d
$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$	Échantillon d'apprentissage *i.i.d.* selon P
$\lvert S \rvert$	Cardinal de l'ensemble S
$(P)^m$	Distribution de l'échantillon S de taille m
$S \sim (P)^m$	S est de taille m, ses éléments sont tirés *i.i.d.* selon P
$(\mathbf{x}, y) \sim P$	$(\mathbf{x}, y)$ est tiré *i.i.d.* selon P
$\mathbf{x} \sim D$	$\mathbf{x}$ est tiré *i.i.d.* selon D
$\mathcal{H}$	Classe d'hypothèses
π	Le prior : distribution *a priori* sur $\mathcal{H}$
ρ	Le posterior : distribution *a posteriori* sur $\mathcal{H}$
$\mathrm{KL}(\rho \Vert \pi)$	Divergence de Kullback-Leibler entre les distributions ρ et π
$G_\rho(\cdot)$	Classifieur de Gibbs
$B_\rho(\cdot)$	Vote de majorité ρ-pondéré sur $\mathcal{H}$
$\mathbf{I}(a)$	Fonction indicatrice : renvoie 1 si a est vraie, 0 sinon
$\mathrm{sign}(a)$	Retourne le signe de a : renvoie 1 si $a \geq 0$, -1 sinon
$\mathbf{v}$	Un vecteur arbitraire
$\mathbf{M}$	Une matrice arbitraire
$\mathrm{diag}(\mathbf{v})$	Matrice diagonale de coefficients $\mathbf{v}$
$\mathbf{0}$	Matrice nulle
$\langle \cdot, \cdot \rangle$	Produit scalaire entre deux vecteurs
$\Vert \cdot \Vert_1$	La norme 1
$\Vert \cdot \Vert_2$	La norme 2
$\Vert \cdot \Vert_\infty$	La norme infinie
$\Vert \cdot \Vert_{op}$	La norme operateur
$\mathbf{Pr}\,(\cdot)$	Probabilité d'un événement
$\mathbf{E}\,(\cdot)$	Espérance d'une variable aléatoire
$\mathbf{Var}\,(\cdot)$	Variance d'une variable aléatoire
Π	Le nombre Pi

Première partie

Préliminaires

Apprentissage supervisé 1

Ce chapitre décrit le contexte de l'apprentissage automatique dans lequel les travaux présentés dans ce mémoire se placent. Comme introduit précédemment, la problématique de l'apprentissage automatique peut se présenter simplement de la manière suivante : l'objectif est d'apprendre une fonction capable d'expliciter au mieux la relation entre un espace d'entrée, par exemple un espace de représentation des e-mails, et un espace de sortie, par exemple une classe "spam" ou "ham". Étant donnés un échantillon de données observées représentatif de la tâche, appelé échantillon d'apprentissage, et un espace d'hypothèses $\mathcal{H}$, l'apprenant doit trouver l'hypothèse de l'espace $\mathcal{H}$ capable d'approximer au mieux la relation existante entre l'espace d'entrée et l'espace de sortie pour le problème considéré. L'idée est donc de chercher une hypothèse consistante avec les données d'apprentissage et suffisamment générale pour se comporter correctement sur les données non vues pendant la phase d'apprentissage. Historiquement, une des premières formalisations de la notion d'induction a été proposée par Tom Mitchell en 1978 avec l'espace des versions [1]. De manière intuitive, l'idée est d'organiser l'espace d'hypothèses sous la forme

1. Le lecteur peut se référer à [Mitchell, 1982] pour la théorie et de l'algorithmie concernant l'espace des versions.

d'un treillis défini en fonction d'une relation d'ordre donnée. L'apprentissage est alors vu comme un processus itératif visant à parcourir efficacement l'espace de recherche afin de trouver une hypothèse consistante avec les données d'apprentissage. Ce type de paradigme s'avère très intéressant lorsque l'espace d'hypothèses est fini et que les hypothèses sont structurées, comme par exemple l'apprentissage de règles pour des systèmes experts, de programmes logiques en programmation logique inductive, ou de grammaires en inférence grammaticale. Cependant, il se montre moins naturel dès lors que les données à traiter sont numériques, comme c'est le cas en multimédia ou en traitement d'images où l'espace d'hypothèses correspond souvent à un ensemble d'hyperplans séparateurs dans un espace vectoriel réel de dimension d. Cet ensemble est souvent infini et donc plus difficile à parcourir de manière structurée. Dans ce contexte, on a généralement recours à des approches par optimisation mathématique. Dans une telle situation, se pose également le problème de mesurer la capacité en généralisation d'une hypothèse. Pour répondre, entre autre, à ces problématiques, de nouveaux modèles basés sur une modélisation statistique de l'apprentissage ont ensuite été proposés et se sont peu à peu imposés. Ces modèles définissent l'apprentissage statistique, dont les fondements ont été introduits par Vladimir Vapnik dans les années 70, puis par Leslie Valiant dans les années 80. C'est dans ce cadre que se positionne cette thèse. En particulier, nous nous intéressons à l'apprentissage dit supervisé [2] que nous présentons dans la suite.

Tout d'abord, dans la section 1.1, nous introduisons plus formellement la tâche de classification supervisée, ainsi que quelques notations [3]. Ensuite, nous énonçons les stratégies classiques pour résoudre cette tâche en section 1.2. Nous détaillons, dans la section 1.3, quatre méthodes classiques permettant de dériver des capacités en généralisation. Enfin, nous présentons trois algorithmes d'apprentissage en section 1.4.

1.1 Un peu de formalisme

Soit X l'espace d'entrée ou de description des données et Y l'espace de sortie ou l'ensemble d'étiquetage. Nous considérons uniquement le cas où les données sont décrites par des vecteurs à valeurs réelles de dimension finie d, c'est-à-dire $X \subseteq \mathbb{R}^d$. Il est important de souligner que rien ne nous empêche de considérer des données plus complexes dites données structurées. L'espace Y peut lui aussi prendre différente formes :

- lorsque Y est un ensemble continu (par exemple $Y = [-1, +1]$ ou $Y = \mathbb{R}$), l'objectif est d'apprendre une quantité, un score, un classement, ... On parle alors de régression ;
- lorsque Y est un ensemble discret, l'objectif est d'apprendre une étiquette, une classe, ... On parle alors de classification. Dans ce mémoire, nous distinguons

2. Notons qu'il existe d'autres paradigmes d'apprentissage tels que l'apprentissage non-supervisé, l'apprentissage semi-supervisé, l'apprentissage par renforcement, l'apprentissage par transfert.

3. Un tableau récapitulatif des principales notations est disponible page 9.

la classification binaire (avec $Y = \{-1, 1\}$) de la classification multiclasse (avec $Y = \{1, \ldots, Q\}$ où $Q > 2$ est un nombre fini de classes).

Étant donnée P une distribution de probabilité jointe (fixée et inconnue) sur $X \times Y$, que l'on appellera un domaine, les ingrédients de l'apprentissage supervisé sont :

- un échantillon d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ de taille m constitué de m observations indépendemment et identiquement distribuées (*i.i.d*) selon P. On note $(P)^m$ la distribution de S et D la distribution marginale de X associée ;
- une classe d'hypothèses $\mathcal{H}$ qui est un ensemble de fonctions $h : X \mapsto Y$ appelées indistinctement hypothèses, modèles, votants, mais aussi classifieurs ou classificateurs en classification, ou régresseurs en régression.

À l'aide de l'information portée par l'échantillon S, l'apprenant doit choisir une hypothèse h dans $\mathcal{H}$. Cette hypothèse h doit décrire au mieux la relation qui existe entre les espaces X et Y. Autrement dit, h doit avoir une bonne qualité en généralisation : étant donné un nouvel exemple $(\mathbf{x}, y)$ tiré aléatoirement selon P, la prédiction $h(\mathbf{x})$ de l'hypothèse doit être la plus proche de la vraie valeur y. Pour mesurer cette qualité on fait généralement appel à une fonction de perte $\ell : Y \times Y \mapsto [0, 1]$ qui associe un coût $\ell(h(\mathbf{x}), y)$ à la prédiction de h sur un exemple $(\mathbf{x}, y)$. Cette notion de risque sert à évaluer une mauvaise réponse relativement à la fonction de perte. Nous définissons maintenant la notion de risque réel.

Définition 1.1 (Risque réel) *Étant donnée une fonction de perte $\ell : Y \times Y \mapsto [0, 1]$, le risque réel (ou l'erreur en généralisation) $\mathbf{R}_P^\ell(h)$ d'une hypothèse h issue de $\mathcal{H}$ sur un domaine P est défini comme l'espérance de la fonction de perte sur le domaine P :*

$$\mathbf{R}_P^\ell(h) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \ell(h(\mathbf{x}), y).$$

On définit le meilleur modèle h^* issu de $\mathcal{H}$ comme le minimiseur du risque réel. Cependant, quelle que soit l'hypothèse $h \in \mathcal{H}$ considérée, $\mathbf{R}_P^\ell(h)$ ne peut être calculé puisque que la loi de probabilité P sur $X \times Y$ est inconnue. Les seules informations disponibles sont, en fait, celles portées par l'échantillon d'apprentissage S. Nous définissons alors le risque empirique évalué sur S.

Définition 1.2 (Risque empirique) *Étant donnés une fonction de perte $\ell : Y \times Y \mapsto [0, 1]$ et un échantillon d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ i.i.d. selon un domaine P, le risque empirique $\mathbf{R}_S(h)$ d'une hypothèse h issue de $\mathcal{H}$ évalué sur S est défini comme l'espérance de la perte estimée sur l'ensemble S :*

$$\mathbf{R}_S^\ell(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, (\mathbf{x}_i, y_i)).$$

Nous aurons aussi besoin de la définition suivante mesurant le désaccord entre deux hypothèses.

Définition 1.3 (Désaccord réel et empirique) *Étant donnée une fonction de perte $\ell : Y \times Y \mapsto [0,1]$ et une distribution D sur X, le désaccord réel $\mathbf{R}^{\ell}_{D}(h,h')$ entre deux hypothèses h et h' issues de $\mathcal{H}$ sur D est défini par :*

$$\mathbf{R}^{\ell}_{D}(h,h') = \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \ell(h(\mathbf{x}),h'(\mathbf{x})). \tag{1.1}$$

Étant donné un échantillon d'exemples non étiquetés $S_u = \{\mathbf{x}_i\}_{i=1}^{m}$ i.i.d. selon D, le désaccord empirique $\mathbf{R}_{S_u}(h,h')$ entre deux hypothèses h et h' issues de $\mathcal{H}$ évalué sur S_u est défini par :

$$\mathbf{R}^{\ell}_{S_u}(h,h') = \frac{1}{m}\sum_{i=1}^{m} \ell(h(\mathbf{x}_i),h'(\mathbf{x}_i)). \tag{1.2}$$

La mesure de risque la plus naturelle revient à comptabiliser les erreurs commises par h sur le domaine P. Dans le cas d'un classifieur, on utilise la fonction de perte $0-1$, $\ell_{0\text{–}1} : Y \times Y \mapsto \{0,1\}$, définie pour un exemple $(\mathbf{x},y)$ par :

$$\begin{aligned}\ell_{0\text{–}1}(h(\mathbf{x}),y) &= \mathbf{I}(h(\mathbf{x}) \neq y)\\ &= \begin{cases} 1 \text{ si } h(\mathbf{x}) \neq y \\ 0 \text{ sinon.}\end{cases}\end{aligned}$$

Par abus de notations, nous posons :

$$\begin{aligned}\mathbf{R}_S(h) = \mathbf{R}^{\ell_{0\text{–}1}}_{S}(h) \quad &; \quad \mathbf{R}_P(h) = \mathbf{R}^{\ell_{0\text{–}1}}_{P}(h)\\ \mathbf{R}_S(h,h') = \mathbf{R}^{\ell_{0\text{–}1}}_{S}(h,h') \quad &; \quad \mathbf{R}_D(h,h') = \mathbf{R}^{\ell_{0\text{–}1}}_{D}(h,h').\end{aligned}$$

Tout au long de ce manuscrit, $\mathbf{R}_S(h)$ et $\mathbf{R}_P(h)$ seront respectivement appelés erreur empirique et erreur réelle. L'extension de la perte $0-1$ à des hypothèses à valeurs réelles est la perte linéaire, $\ell_{lin} : \mathbb{R} \times \mathbb{R} \mapsto [0,1]$, qui reste intéressante en classification binaire et qui est définie pour un exemple $(\mathbf{x},y)$ par :

$$\ell_{lin}(h(\mathbf{x}),y) = \frac{1}{2}\left(1 - yh(\mathbf{x})\right).$$

Comme nous allons le voir dans la section suivante, une stratégie pour trouver la meilleure hypothèse au sens de l'erreur réelle revient, en partie, à minimiser l'erreur empirique. Cependant minimiser empiriquement la valeur de la perte $0-1$ est un problème NP-difficile. Il est donc généralement nécessaire de faire appel à une relaxation convexe de la perte $0-1$ (on parle souvent de *surrogate loss*). Parmi ces relaxations, nous pouvons citer la perte hinge qui est connue pour être la meilleure approximation de la perte $0-1$ [Ben-David *et al.*, 2012a] et que nous utiliserons à plusieurs reprises dans ce manuscrit. Pour un exemple $(\mathbf{x},y)$, la perte hinge est définie par :

$$\begin{aligned}\ell_{hinge}(h(\mathbf{x}),y) &= \left[1 - yh(\mathbf{x})\right]_+\\ &= \max(0, 1 - yh(\mathbf{x})).\end{aligned}$$

Les trois fonction de pertes présentées ci-dessus sont représentées sur la figure 1.1.

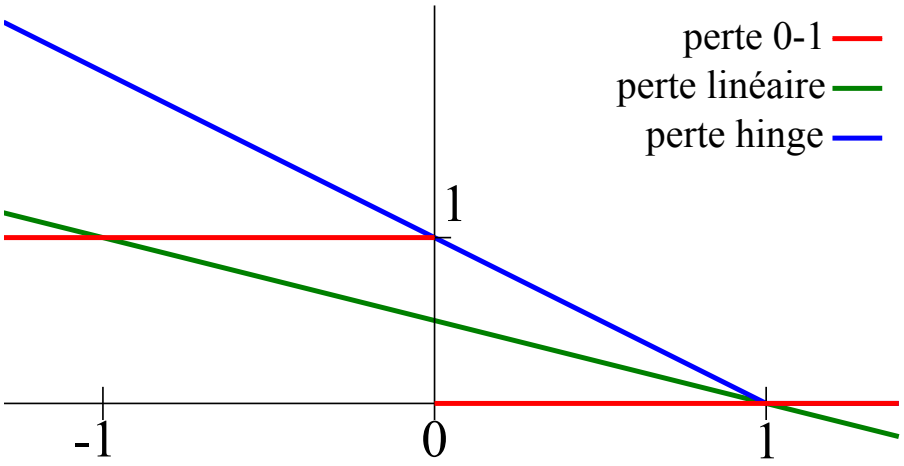


FIGURE 1.1 – *Représentation de la fonction de perte* $0-1$ *(en rouge, la plus claire), la fonction de perte linéaire (en vert, l'intermédiaire) et la fonction de perte hinge (en bleu, la plus foncée).*

1.2 Stratégies classiques de minimisation du risque

Nous nous intéressons maintenant aux stratégies classiques permettant d'apprendre une bonne hypothèse, c'est-à-dire une hypothèse amenant à une erreur réelle faible.

Dans le meilleur des mondes, nous disposerions d'une infinité d'exemples d'apprentissage et minimiser le risque empirique serait la stratégie la plus pertinente. Dans la réalité, la quantité d'exemples disponibles est limitée et il existe souvent une hypothèse h (parfois complexe) d'erreur empirique nulle. Face à de nouvelles données jamais observées, cette hypothèse, "parfaite" en apparence, ne montrera pas toujours de bonnes performances et l'erreur réelle pourra en fait s'avèrer — beaucoup — plus élevée que l'erreur empirique : on parle alors de sur-apprentissage.

Pour éviter ce problème d'apprentissage "par cœur", une solution envisagée est de considérer le compromis biais/variance, se résumant principalement en un juste équilibre entre l'erreur empirique et la complexité de la classe d'hypothèses : selon le principe du rasoir d'Occam, à performance égale, on préfère un modèle simple à un modèle complexe. Nous présentons ci-après trois stratégies classiques permettant de contrôler ce compromis pour apprendre une hypothèse d'erreur réelle faible, puis en section 1.3 nous étudierons la capacité en généralisation de telles hypothèses.

1.2.1 Minimisation du Risque Empirique (ERM)

L'approche la plus intuitive est celle de la minimisation du risque empirique en restreignant la classe d'hypothèses $\mathcal{H}$. Puis, étant donné un échantillon d'apprentissage S, on sélectionne dans $\mathcal{H}$ le minimiseur h_S^* du risque empirique évalué sur S :

$$h_S^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \, \mathbf{R}_S^\ell(h).$$

Si $\mathcal{H}$ est suffisamment restreint, alors le sur-apprentissage sera évité. Cependant, sans informations *a priori* sur la tâche considérée, $\mathcal{H}$ sera difficile à définir et choisir un modèle performant sur les données sera fastidieux.

1.2.2 Minimisation Structurelle du Risque (SRM)

Puisque restreindre $\mathcal{H}$ n'est pas aisé, une solution est de faire appel à un ensemble d'hypothèses ordonnées en fonctions de leur complexité. Concrètement, la minimisation structurelle du risque [Vapnik, 1995] considère un ensemble de familles d'hypothèses $\mathcal{H}_1$, $\mathcal{H}_2$, ... de complexités croissantes telles que : $\forall j \in \{1, 2, \ldots\}$, $\mathcal{H}_i \subset \mathcal{H}_{i+1}$. Dans chacune des familles, on choisit l'hypothèse qui minimise le risque empirique évalué sur S. Enfin, parmi ces hypothèses (sous-)optimales, on sélectionne l'hypothèse h_S^* qui minimise la somme du risque et d'une pénalisation. Plus formellement :

$$h_S^* = \underset{h \in \mathcal{H}_j,\ j=\{1,2,\ldots\}}{\operatorname{argmin}} \left\{ \mathbf{R}_S^\ell(h) + \operatorname{pen}(\mathcal{H}_j) \right\},$$

où $\operatorname{pen}(\mathcal{H}_j)$ pénalise $\mathcal{H}_j$ en fonction de sa complexité.

1.2.3 Minimisation Régularisée du Risque (RRM)

Étant donnée une classe d'hypothèses $\mathcal{H}$, une approche similaire consiste à régulariser le choix de l'hypothèse (selon une norme donnée $\|\cdot\|$). En fait, cette régularisation contrôle la complexité de l'hypothèse : plus sa complexité est élevée, plus elle sera pénalisée. Autrement dit, étant donné l'échantillon S, h_S^* correspond à l'hypothèse de $\mathcal{H}$ minimisant le compromis risque empirique et régularisation :

$$h_S^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left\{ \mathbf{R}_S^\ell(h) + \lambda \|h\| \right\},$$

où λ est l'hyperparamètre de compromis.

1.3 Bornes en généralisation

La mise en œuvre des principes énoncés précédemment doit s'accompagner de garanties théoriques permettant de quantifier la précision de l'estimateur du risque par rapport à sa valeur réelle. La théorie statistique de l'apprentissage de [Vapnik, 1995] étudie les conditions de consistance et donc de convergence de la minimisation du risque empirique vers la valeur réelle. On parle alors de bornes en généralisation, souvent mentionnée en tant que bornes PAC (*Probably Approximately Correct*) [Valiant, 1984] ayant la forme suivante :

$$\underset{S \sim (P)^m}{\mathbf{Pr}} \left\{ \left| \mathbf{R}_P^\ell(h) - \mathbf{R}_S^\ell(h) \right| \leq \epsilon \right\} \geq 1 - \delta,$$

où $\epsilon \geq 0$ et $\delta \in (0, 1]$. En d'autre termes, on veut minorer la probabilité que l'écart entre le risque réel et son estimation soit inférieur à un certain ϵ, le plus faible possible. Plus cet ϵ est faible, plus l'estimation est précise. La question majeure consiste donc à savoir si minimiser $\mathbf{R}_S^\ell(\cdot)$ avec un échantillon de taille infinie permet de minimiser $\mathbf{R}_P^\ell(\cdot)$. Bien entendu, cette stratégie doit être réalisable en pratique, c'est-à-dire lorsque l'échantillon S est fini, impliquant qu'il doit y avoir convergence de $\mathbf{R}_S^\ell(\cdot)$ vers $\mathbf{R}_P^\ell(\cdot)$.

Plus l'écart entre le risque empirique et le risque réel est faible, plus les garanties sont élevées et plus l'estimateur est précis. La dérivation de telles bornes réside principalement dans l'utilisation d'inégalités de concentration [4]. Par exemple, la contribution du chapitre 5 se base sur ce principe.
Nous développons rapidement quatre approches permettant d'obtenir de telles bornes théoriques. Les deux premières prennent en compte la complexité de la classe d'hypothèses (avec la dimension de Vapnik-Chervonenkis et la complexité de Rademacher). Puis nous présenterons deux méthodes permettant de considérer l'algorithme d'apprentissage utilisé en s'affranchissant de cette complexité. De plus, dans le chapitre 3, nous présenterons plus en détails une cinquième approche : la théorie PAC-Bayésienne.

1.3.1 Convergence uniforme

Une des mesures de base originelle est la dimension de Vapnik-Chervonenkis [Vapnik et Chervonenkis, 1971, Vapnik, 1982] (notée VC-dim.) qui permet de mesurer la complexité d'une classe d'hypothèses $\mathcal{H}$ en classification binaire.

Définition 1.4 *La dimension de Vapnik-Chervonenkis* $\mathrm{VC}(\mathcal{H})$ *d'une classe d'hypothèses* $\mathcal{H}$ *en classification binaire est définie comme étant le cardinal maximal d'un sous-ensemble* $X' \subset X$ *tel qu'on puisse toujours trouver une fonction* h *dans* $\mathcal{H}$ *qui classe parfaitement tous les élément de* X'*, quelle que soit leur étiquette. Plus formellement :*

$$\mathrm{VC}(\mathcal{H}) = \max\left\{ \left|X'\right| : \forall y_i \in \{-1,+1\}^{|X'|},\ \exists h \in \mathcal{H},\ \text{telle que } \forall \mathbf{x}_i \in X',\ h(\mathbf{x}_i) = y_i \right\}.$$

La VC-dim. correspond à la quantité maximale de données telle qu'il existe une hypothèse dans $\mathcal{H}$ qui soit consistante avec n'importe quel étiquetage. Cette définition permet alors de majorer l'écart maximal entre $\mathbf{R}_P^\ell(\cdot)$ et $\mathbf{R}_S^\ell(\cdot)$ relativement à la fonction de perte $\ell(\cdot,\cdot)$ considérée, de la taille m de l'échantillon d'apprentissage et de la complexité de la classe d'hypothèses $\mathcal{H}$ mesurée par la VC-dim.

Théorème 1.1 *Soit* X *un espace d'entrée, soit* $Y = \{-1,+1\}$ *l'ensemble de classes et soit* P *un domaine sur* $X \times Y$*. Soit* S *un échantillon de* m *exemples tirés i.i.d. selon* P*, soit* $\mathcal{H}$ *un ensemble continu d'hypothèses de* X *vers* Y *avec une VC-dim.* $\mathrm{VC}(\mathcal{H})$*. Alors, pour tout* $\delta \in (0,1]$*, avec une probabilité d'au moins* $1-\delta$ *sur le choix de* $S \sim (P)^m$*, on a :*

$$\sup_{h\in\mathcal{H}} \left| \mathbf{R}_P^\ell(h) - \mathbf{R}_S^\ell(h) \right| \ \leq\ \sqrt{\frac{\mathrm{VC}(\mathcal{H})\left(\ln\frac{2m}{\mathrm{VC}(\mathcal{H})}+1\right)+\ln\left(\frac{4}{\delta}\right)}{m}}.$$

L'utilisation du *supremum* sur $\mathcal{H}$ rend ce théorème difficile à manipuler. Cependant, nous pouvons directement en dériver une borne non-asymptotique plus simple à utiliser et valable pour toute hypothèse de $\mathcal{H}$ et en particulier pour l'hypothèse qui minimise le risque empirique.

4. Voir [Boucheron *et al.*, 2004] pour plus d'informations concernant les inégalités de concentration.

Corollaire 1.1 *Sous les mêmes hypothèses que le théorème 1.1 précédent, avec une probabilité d'au moins* $1-\delta$ *sur le choix de l'échantillon aléatoire* $S \sim (P)^m$*, on a :*

$$\forall h \in \mathcal{H},\ \mathbf{R}_P^\ell(h) \ \leq\ \mathbf{R}_S^\ell(h) + \sqrt{\frac{\mathrm{VC}(\mathcal{H})\left(\ln \frac{2m}{\mathrm{VC}(\mathcal{H})}+1\right)+\ln\left(\frac{4}{\delta}\right)}{m}}.$$

Notons que lorsque la VC-dim. de $\mathcal{H}$ est connue, alors la partie droite de la borne est calculable. Cette borne indique qu'avec une confiance de $1-\delta$, le risque empirique d'une hypothèse tend vers sa valeur réelle lorsque la taille m de l'échantillon d'apprentissage augmente, et ce, d'autant plus "vite" que la VC-dim. de $\mathcal{H}$ est faible. Quand la VC-dim. de $\mathcal{H}$ est finie, si h_S^* est l'hypothèse qui minimise le risque empirique, alors la loi des grands nombres et la convergence uniforme impliquent que son risque réel et son risque empirique convergent tous les deux en probabilité vers le minimum du risque sur $\mathcal{H}$. Lorsqu'un algorithme d'apprentissage permet de vérifier cette propriété, on dit qu'il est consistant.

En pratique, les arguments portées par cette théorie peuvent poser des inconvénients majeurs. Tout d'abord, l'obtention d'une nouvelle borne basée sur la VC-dim. est en général fastidieux. D'une part, la VC-dim. est parfois complexe à calculer, d'autre part, il existe des classes d'hypothèses de VC-dim. infinie (c'est le cas des k plus proches voisins présentés en section 1.4). De plus, c'est une borne peu précise et qui montre en pratique peu d'utilité : il est généralement nécessaire de disposer d'une quantité considérable de données pour obtenir des résultats significatifs, précis et pertinents. En effet, il n'est pas rare d'avoir une borne supérieure à 1 même lorsque l'on dispose de milliers d'exemples. Enfin, cette analyse ne prend ni en compte l'algorithme utilisé, ni la distribution des données. Elle peut être vue en ce sens comme une analyse "dans le pire cas".

La complexité de Rademacher, présentée dans la section suivante, permet, en partie, de contourner cette difficulté. En effet, alors que la VC-dim. se focalise sur le pire étiquetage pour $\mathcal{H}$, la complexité de Rademacher se calcule en moyenne sur tous les étiquetages possibles et permet d'introduire de l'information sur la distribution des données.

1.3.2 Complexité de Rademacher

Intuitivement, la complexité de Rademacher[5] mesure la capacité d'une classe d'hypothèses à résister au bruit et peut amener à des bornes plus précises que celles basées sur la VC-dim. [Koltchinskii et Panchenko, 1999]. Elle tire son nom de l'utilisation de variables de Rademacher définies ci-dessous.

Définition 1.5 (Variable de Rademacher) *Une variable de Rademacher* κ *est une variable aléatoire*

5. Il existe une mesure comparable appelée la complexité de Gauss [Bartlett et Mendelson, 2002] faisant appel à des variables gaussiennes.

telle que :

$$\kappa = \begin{cases} +1 \text{ avec une probabilité } \frac{1}{2} \\ -1 \text{ sinon.} \end{cases}$$

Une variable de Rademacher modélise donc un étiquetage binaire aléatoire. Ainsi, la complexité de Rademacher empirique, calculée sur un échantillon (non-étiqueté) de taille m, se définit par :

Définition 1.6 (Complexité empirique et réelle de Rademacher) *Soit $S = \{\mathbf{x}_i\}_{i=1}^m$ un échantillon (non-étiqueté). Soit $\mathcal{H}$ une classe d'hypothèses. La complexité empirique de Rademacher évaluée sur S de $\mathcal{H}$ est :*

$$\mathfrak{R}_S(\mathcal{H}) = \mathop{\mathbf{E}}_{\boldsymbol{\kappa}} \left[\sup_{h \in \mathcal{H}} \left| \frac{2}{m} \sum_{i=1}^{m} \kappa_i h(\mathbf{x}_i) \right| \right], \tag{1.3}$$

où $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_m)^\top$ est un vecteur de m variables de Rademacher indépendantes.
La complexité de Rademacher d'une classe d'hypothèses $\mathcal{H}$ est alors définie par l'espérance de $\mathfrak{R}_S(\mathcal{H})$ sur tous les échantillons S de taille m :

$$\mathfrak{R}_m(\mathcal{H}) = \mathop{\mathbf{E}}_{S \sim (P)^m} \mathfrak{R}_S(\mathcal{H}). \tag{1.4}$$

L'équation (1.3), $\mathfrak{R}_S(\mathcal{H})$, se focalise sur un échantillon S, alors que l'équation (1.4), $\mathfrak{R}_m(\mathcal{H})$, est l'espérance de $\mathfrak{R}_S(\mathcal{H})$ sur tous les échantillons de taille m possibles et modélise une information plus riche. Ces deux définitions viennent en opposition au terme $VC(\mathcal{H})$ qui considère l'étiquetage le plus difficile pour $\mathcal{H}$, puisque la complexité de Rademacher calcule une espérance sur tous les étiquetages possibles. Le théorème suivant énonce deux bornes en généralisation basée sur la complexité de Rademacher proposées par [Koltchinskii et Panchenko, 1999, Bartlett et Mendelson, 2002].

Théorème 1.2 *Soit $\mathcal{H}$ une classe d'hypothèses et $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ un échantillon fini d'exemples tirés i.i.d selon un domaine P. Alors, pour tout $\delta \in (0, 1]$, avec une probabilité d'au moins $1 - \delta$ sur le choix de $S \sim (P)^m$, on a :*

$$\forall h \in \mathcal{H},\ \mathbf{R}_P^\ell(h) \ \leq\ \mathbf{R}_S^\ell(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}, \tag{1.5}$$

$$\mathbf{R}_P^\ell(h) \ \leq\ \mathbf{R}_S^\ell(h) + \mathfrak{R}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{1.6}$$

La seconde borne (1.6) dépend de la complexité de Rademacher empirique $\mathfrak{R}_S(\mathcal{H})$ qui peut facilement être calculée à partir de l'échantillon S et permet, de plus, d'introduire une information sur la distribution des données. Lorsque $\mathfrak{R}_S(\mathcal{H})$ est calculable, la borne peut être informative et plus précise que les bornes de type VC-dim. Cependant, ce calcul est parfois complexe, voir impossible, ce qui rend difficile sa manipulation.

Pour contourner ce problème, nous allons voir maintenant comment la prise en compte de l'exploration de la classe d'hypothèses par l'algorithme permet de s'affranchir du terme de complexité lié à la classe d'hypothèses.

1.3.3 Stabilité uniforme

La manière dont l'espace d'hypothèses est exploré dépend de l'algorithme (comme par exemple les approches régularisées du type RRM, section 1.2.3). Il s'avère donc important de relier les capacités en généralisation aux spécificités de l'algorithme. Les résultats principaux de [Bousquet et Elisseeff, 2002] ont pour objectif de tirer partie de ces spécificités via la notion suivante de stabilité algorithmique uniforme, pouvant amener à des bornes précises. Intuitivement, un algorithme est stable s'il est robuste à de faibles modifications de l'échantillon d'apprentissage considéré en entrée de l'algorithme. Autrement dit, sa sortie ne change pas significativement en fonction de ces modifications.

Définition 1.7 (Stabilité uniforme [Bousquet et Elisseeff, 2002]) *Un algorithme $\mathcal{A}$ admet une stabilité uniforme β au regard de la fonction de perte $\ell(\cdot,\cdot)$ utilisée par $\mathcal{A}$ s'il vérifie :*

$$\forall S \in (X \times Y)^m,\ \forall i \in \{1,\dots,m\},\ \sup_{(\mathbf{x},y)\in S} \left| \ell\left(h_S(\mathbf{x}),y\right) - \ell\left(h_{S^{\setminus i}}(\mathbf{x}),y\right) \right| \leq \beta,$$

où h_S est le modèle appris par $\mathcal{A}$ depuis S et $h_{S^{\setminus i}}$ celui appris par $\mathcal{A}$ depuis S privé de l'exemple $(\mathbf{x}_i, y_i)$.

Le terme β est relié à la fonction de perte et à la régularisation utilisée par l'algorithme. Lorsqu'il est vu comme une fonction de m et qu'il décroît en $\frac{1}{m}$, on parle d'algorithme stable uniformément. [Bousquet et Elisseeff, 2002] ont démontré que de nombreux algorithmes de type RRM vérifient cette définition. Lorsqu'elle est satisfaite, ils ont prouvé la borne en généralisation suivante qui offre une borne de consistance précise lorsque β est de l'ordre de $\frac{1}{m}$.

Théorème 1.3 ([Bousquet et Elisseeff, 2002]) *Soit $\mathcal{A}$ un algorithme de stabilité uniforme β au regard de la fonction de perte $\ell(\cdot,\cdot)$ utilisée par $\mathcal{A}$ telle que $0 \leq \ell\left(h_S(\mathbf{x}),y\right) \leq \ell^{UP}$ pour tout $(\mathbf{x},y) \in X \times Y$ et tout ensemble S. Alors, pour tout $m \geq 1$ et pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$, on a :*

$$\mathbf{R}_P^{\ell}\left(h_S\right) \ \leq\ \mathbf{R}_S^{\ell}\left(h_S\right) + 2\beta + \left(4m\beta + \ell^{UP}\right)\sqrt{\frac{\ln\frac{1}{\delta}}{2m}},$$

où h_S est le modèle appris par $\mathcal{A}$ depuis S.

Contrairement à la notion de convergence uniforme, la stabilité uniforme permet, d'une part, de prendre en compte la régularisation et, d'autre part, de s'affranchir du terme explicite lié à la complexité de la classe d'hypothèses. Notons que β prend implicitement en compte cette complexité. Des bornes consistantes peuvent donc être dérivées lorsque la complexité est difficile à analyser ou vaut l'infini.

Cependant, [Xu et Mannor, 2010, Xu et Mannor, 2012, Xu *et al.*, 2012] ont démontré que les algorithmes construisant des modèles parcimonieux (par exemple à l'aide d'une régularisation de type $\|\cdot\|_1$) ne sont pas stables face à une faible modification de l'échantillon. Il s'agit de cas où la parcimonie peut être vue comme une capacité à

identifier des attributs redondants. Pour contrer ce problème, ils ont proposé la notion suivante de robustesse algorithmique, dont on tirera bénéfice dans le chapitre 6.

1.3.4 Robustesse algorithmique

Intuitivement, un algorithme est dit robuste [Xu et Mannor, 2010, Xu et Mannor, 2012] au sens d'une modification de l'échantillon d'apprentissage S, s'il renvoie des performances similaires sur S et sur un échantillon test T "proche" de S : *"if a testing sample is similar to a training sample then the testing error is close to the training error"*. Cette notion de similarité est capturée à l'aide d'un partitionnement de l'espace $X \times Y$ construit de telle sorte que deux exemples proches et de même classe appartiennent à la même partition. La définition de cette partition repose sur la notion de nombre de couvertures [Kolmogorov et Tikhomirov, 1959] :

Définition 1.8 (Nombre de couvertures) *Soit un espace métrique (Z, ϱ) où $\varrho(\cdot)$ est une métrique sur Z, soit $Z' \subset Z$. On dit que $\widehat{Z}' \subset Z'$ est une γ-courverture de Z' si pour tout élément t de Z', il existe un élément $\widehat{t}$ dans $\widehat{Z}'$ tel que : $\varrho(t, \widehat{t}) \leq \gamma$. Le nombre de γ-couvertures de Z' est alors :*

$$\mathcal{N}(\gamma, Z', \varrho) = \min\left\{|\widehat{Z}'| : \widehat{Z}' \textit{ est une } \gamma\textit{-couverture de } Z'\right\}.$$

En particulier, si X est un espace compact, son nombre de γ-couvertures $\mathcal{N}(\gamma, X, \varrho)$ est fini. Dans ce cas, le nombre de γ-couvertures de $X \times Y$ est fini et vaut[6] $|Y|\mathcal{N}(\gamma, X, \varrho)$. En effet, ces γ-couvertures sont définies de telle sorte que si deux exemples $(\mathbf{x}, y)$ et $(\mathbf{x}', y')$ appartiennent au même sous-ensemble, alors ils sont de même classes ($y = y'$) et sont proches selon la métrique $\varrho(\cdot, \cdot)$ ($\varrho(\mathbf{x}, \mathbf{x}') \leq \gamma$). Ainsi, cette formule de partitionnement permet de définir un algorithme robuste de la manière suivante.

Définition 1.9 (Algorithme robuste [Xu et Mannor, 2010, Xu et Mannor, 2012]) *Soit un échantillon d'apprentissage S constitué de m exemples tirés i.i.d. selon un domaine P sur $X \times Y$. Un algorithme $\mathcal{A}$ est $(M, \boldsymbol{\epsilon}(S))$ robuste sur P au regard de sa fonction de perte $\ell(\cdot, \cdot)$, avec $M \in \mathbb{N}$ et $\epsilon : (X \times Y)^m \mapsto \mathbb{R}$, si $X \times Y$ peut être partitionné en M ensembles disjoints, notés $\{Z_j\}_{j=1}^{M}$, tels que pour tout exemple $(\mathbf{x}, y)$ appartenant à S, pour tout $(\mathbf{x}', y')$ tiré selon le domaine P et pour tout $j \in \{1, \ldots, M\}$, on a :*

$$\big((\mathbf{x}, y), (\mathbf{x}', y')\big) \in Z_j^2 \quad \Longrightarrow \quad \big|\ell(h_S(\mathbf{x}), y) - \ell(h_S(\mathbf{x}'), y')\big| \leq \epsilon(S),$$

où h_S est le modèle appris par $\mathcal{A}$ depuis S.

Étant donné un échantillon d'apprentissage S, la robustesse d'un algorithme, mesurée par les valeurs de M et $\epsilon(S)$, dépend donc de S. Lorsque cette définition n'est pas vérifiée pour tous les exemples de S, les auteurs ont proposé une relaxation pour laquelle la condition peut uniquement être vérifiée sur un sous-ensemble de S (on parle alors de *pseudo-robustesse*[7]).

6. On calcule la γ-couverture sur X pour chaque classe de Y.
7. La définition de la pseudo-robustesse est donnée dans [Xu et Mannor, 2010, Xu et Mannor, 2012].

Un algorithme vérifiant la définition 1.9 montre les garanties en généralisation suivantes.

Théorème 1.4 ([Xu et Mannor, 2010, Xu et Mannor, 2012]) *Si un ensemble d'apprentissage* $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ *est constitué de m exemples tirés i.i.d. selon un domaine P sur* $X \times Y$ *et si l'algorithme* $\mathcal{A}$ *est* $(M, \epsilon(S))$ *robuste sur P au regard de la fonction de perte* $\ell(\cdot,\cdot)$*, telle que* $0 \leq \ell\left(h_S(\mathbf{x}), y\right) \leq \ell^{UP}$*, alors pour tout* $\delta \in (0,1]$*, avec une probabilité d'au moins* $1-\delta$ *sur le choix aléatoire de* $S \sim (P)^m$*, on a :*

$$\mathbf{R}_P^\ell(h_S) \;\leq\; \mathbf{R}_S^\ell(h_S) + \epsilon(S) + \ell^{UP}\sqrt{\frac{2M\ln 2 + 2\ln\frac{1}{\delta}}{m}},$$

où h_S *est le modèle appris par* $\mathcal{A}$ *depuis S.*

Puisque $\epsilon(S)$ dépend de la γ-couverture de $X \times Y$ et donc de sa taille M, il existe un compromis entre M et $\epsilon(S)$. Similairement à la capacité en généralisation d'un algorithme stable, le terme de complexité de la classe d'hypothèses n'intervient pas et permet d'obtenir des garanties de consistance même lorsque la complexité ne peut être calculée. Alors que la stabilité étudie la variation du coût renvoyé par la fonction de perte pour de faibles changements de l'échantillon d'apprentissage et suggère que l'hypothèse apprise ne varie que très peu, la robustesse se focalise sur la divergence entre les coûts associées à deux exemples proches et implique que l'hypothèse apprise doit être localement consistante.
Cette notion d'algorithme robuste permet, entre autres, de considérer des contextes d'apprentissage non standard, tels que les chaînes de Markov [Xu et Mannor, 2010, Xu et Mannor, 2012], ou le problème de l'adaptation de domaine que nous présenterons dans le chapitre 2. Ceci nous permettra de dériver une borne en généralisation pour l'algorithme d'adaptation de domaine que nous présenterons dans le chapitre 6.

Nous énonçons, par la suite, trois méthodes d'apprentissage supervisé auxquelles nous ferons appel dans ce manuscrit.

1.4 Quelques méthodes de classification supervisée

Nous présentons dans cette section trois méthodes d'apprentissage supervisé. Nous commençons par la plus intuitive en section 1.4.1 : les k Plus Proches Voisins (k-PPV ou k-NN). Un classifieur de type k-PPV étiquette un exemple en renvoyant la classe majoritaire dans son k-voisinage. La contribution du chapitre 4 s'intéresse à améliorer cette approche. Nous rappelons ensuite un des algorithmes les plus populaires en apprentissage automatique : les Machines à Vecteurs de Support (SVM) que nous considérerons comme modèle de référence tout au long de ce mémoire. Enfin, nous énonçons la théorie proposée par [Balcan *et al.*, 2008a, Balcan *et al.*, 2008b] pour apprendre un classifieur linéaire dans un espace de projection défini par une fonction de similarité dite (ϵ,γ,τ)-bonne. Cette théorie servira de base au chapitre 6 pour développer une

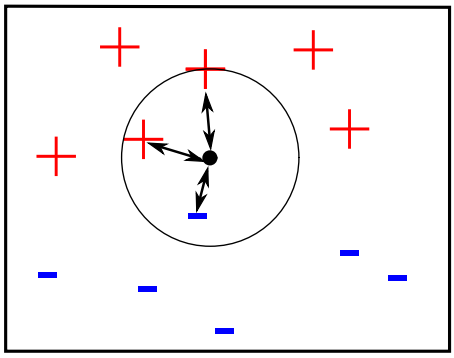


FIGURE 1.2 – *Exemple de classification avec un classifieur* 3*-PPV d'un objet (en noir). Ici, la classe majoritaire dans le* 3*-voisinage est* $+1$.

méthode d'adaptation d'un classifieur d'un domaine vers un autre.
Signalons que la présentation des k-PPV et celle des SVM s'inspirent de celles proposées dans [Cornuéjols et Miclet, 2010].

1.4.1 Le plus intuitif : les k plus proches voisins

Le principe des plus proches voisins est très simple et intuitif [Cover et Hart, 1967, Duda *et al.*, 2001] (comme illustré par la figure 1.2) : la classe d'un nouvel exemple correspond à la classe majoritaire des k éléments lui ressemblant le plus dans l'échantillon d'apprentissage. Pour ce faire, nous disposons d'un ensemble d'apprentissage S, d'une fonction de distance[8] $\mathrm{d} : X \times X \mapsto \mathbb{R}$ permettant de comparer deux exemples, d'un nombre de voisins k et d'une fonction de choix de la classe[9] parmi les k voisins les plus proches selon $\mathrm{d}(\cdot,\cdot)$. Autrement dit, étant donné un exemple à étiqueter, on se concentre uniquement sur l'hypersphère contenant les k voisins. L'algorithme des k-PPV est décrit dans l'algorithme 1 et contrairement à de nombreuses méthodes d'apprentissage automatique, il ne requiert pas de phase d'apprentissage à proprement dit, au sens qu'il n'y a pas de phase d'optimisation mathématique.

Algorithme 1 Algorithme des k plus proches voisins

entrée un échantillon d'apprentissage S, un entier k, une fonction de distance $\mathrm{d}(\cdot,\cdot)$, une fonction de choix de classe $c(\cdot)$, l'objet à classer $\mathbf{x}$
sortie la classe affectée à $\mathbf{x}$
  $Voisins(\mathbf{x}) \leftarrow$ Déterminer les k plus proches voisins (dans S) de $\mathbf{x}$ selon $\mathrm{d}(\cdot,\cdot)$
  $y \leftarrow c\,(Voisins(\mathbf{x}))$ (combinaison des classes des k exemples)
  retourner y

Sous l'hypothèse que les probabilités conditionnelles par classe $P(\mathbf{x}|y)$ sont localement constantes, et correctement estimées par l'échantillon d'apprentissage, la règle de décision des k-PPV est cohérente. En effet, dans une telle situation, la classification par k-PPV revient à choisir la classe la plus probable dans le k-voisinage. Concrètement, soit S un échantillon d'apprentissage de taille m. Étant donné un nouvel exemple $\mathbf{x}$,

8. Classiquement, La fonction $d(\cdot,\cdot)$ est la distance euclidienne.
9. Classiquement, la fonction de choix de classes est un vote de majorité sur k plus proches voisins.

la fonction $d(\cdot,\cdot)$ associée à la valeur de k permet de définir une certaine région de volume V qui contient les k voisins de $\mathbf{x}$ dans S (en pratique cette région est une hypersphère comme sur la figure 1.2). Pour toute classe y de Y, on note alors m_y le nombre d'exemples de S de classe y et k_y le nombre de k-voisins de $\mathbf{x}$ d'étiquette y. On estime donc la probabilité conditionnelle sur S par :

$$\widehat{P}_S(\mathbf{x}|y) = \frac{k_y/m_y}{V}.$$

Il est facilement démontrable que lorque la quantité d'exemples d'apprentissage m augmente, alors la valeur empirique $\widehat{P}_S(\mathbf{x}|y)$ converge vers sa valeur réelle. De plus, si $\frac{m_y}{m} = \widehat{P}_S(y)$ est considéré comme un estimateur de la probabilité *a priori* de la classe y on obtient :

$$k_y = mV\widehat{P}_S(\mathbf{x}|y)\widehat{P}_S(y).$$

Ainsi, la classe majoritaire, c'est-à-dire celle associée au k_y le plus élevé, est celle qui maximise $\widehat{P}_S(\mathbf{x}|y)\widehat{P}_S(y) = \widehat{P}_S(y|\mathbf{x})\widehat{P}_S(\mathbf{x})$ (par application du théorème de Bayes). Cela revient à choisir la classe qui correspond à la règle de décision du maximum *a posteriori* : $\max_{y\in Y}\widehat{P}_S(y|\mathbf{x})$.

Concernant les garanties de cette méthode, la théorie classique des k-PPV stipule que pour tout k, plus la quantité m d'exemples augmente, plus l'erreur des k-PPV tend vers le risque bayésien optimal défini par :

$$\mathbf{R}_{bayes} = \mathbf{E}_{\mathbf{x}\sim D}\left(1 - \max_{y\in Y} P(y|\mathbf{x})\right)P(\mathbf{x}).$$

Dans le cas binaire avec le nombre de classe $|Y| = 2$, on a :

$$\mathbf{R}_{bayes} \leq \mathbf{R}_{k\text{-PPV}} \leq \mathbf{R}_{(k-1)\text{-PPV}} \leq \cdots \leq \mathbf{R}_{1\text{-PPV}} \leq 2\mathbf{R}_{bayes},$$

où $\mathbf{R}_{k\text{-PPV}}$ est l'erreur du k-PPV avec :

$$\mathbf{R}_{k\text{-PPV}} \leq \mathbf{R}_{bayes} + \mathbf{R}_{1\text{-PPV}}\sqrt{\frac{2}{\Pi\, k}},$$

où $\Pi \simeq 3.141\ldots$ est le nombre Pi.
Quel que soit le nombre de classes $|Y| = Q$, on a :

$$\mathbf{R}_{1\text{-PPV}} \leq \mathbf{R}_{bayes} + \left(2 - \frac{Q}{Q-1}\mathbf{R}_{bayes}\right).$$

Ces résultats [Ripley, 2007] confirment, d'une part, que plus k est élevé, plus l'estimation est précise et, d'autre part, que la règle des 1-PPV est asymptotiquement efficace (la moitié de l'information pour classer un nouveau point est portée par son plus proche voisin). Cependant, ces propriétés ne sont vraies qu'asymptotiquement (quand $m \to +\infty$).

Deux principales stratégies pour s'attaquer à ces contraintes ont été proposées dans la littérature. La première repose sur le fait que les probabilités conditionnelles par classe sont supposées localement régulières. Or, plus la dimension de l'espace augmente,

moins cette hypothèse est satisfaite. En conséquence, une solution est d'adapter localement les voisinages. Citons en illustration, les travaux de [Hastie et Tibshirani, 1996] qui font appel à une analyse discriminante linéaire locale pour modifier les voisinages, et ceux de [Nock *et al.*, 2003] qui ont proposé SNN, un k-PPV symétrique : la classe d'un exemple est déterminée par la classe majoritaire parmi les k-voisins ainsi que ceux incluant l'exemple dans leur propre k-voisinage. D'autre part, la performance des PPV étant dépendante de la distance intervenant dans le calcul des voisinages, une seconde stratégie fait appel à l'apprentissage de métriques [10]. Par exemple, LMNN (*Large Margin Nearest Neighbor*) apprend une distance de Mahalanobis qui minimise l'erreur empirique d'apprentissage d'un k-PPV avec une marge minimale [Weinberger et Saul, 2009]. Quelle que soit la stratégie, la règle de décision se base sur ces voisinages locaux pouvant aboutir à des phénomènes de sur-apprentissage (notamment en grande dimension). En outre, en pratique, m est limité et k doit être choisi avec précaution. En effet, deux principes s'opposent. D'une part, plus k est grand, plus l'estimation de la densité est pertinente. D'autre part, seuls les voisinages les plus proches (correspondant à un petit k) amènent à une règle de classification plus fiable. Dans la littérature, différentes études théoriques et empiriques ont été menées pour analyser ce compromis entre grande et petite valeur de k, on trouve souvent la valeur $k = \sqrt{\frac{m}{Q}}$.

Pour contrer ces inconvénients, nous proposerons dans le chapitre 4, une approche originale de la classification par k-PPV dont le but est de construire un vote de majorité pondéré sur un ensemble de classifieurs k-PPV.

1.4.2 Un des modèles de référence : les machines à vecteurs de support

Les Machines à Vecteurs de Support [11] (ou Séparateurs à Vaste Marge, SVM) [Boser *et al.*, 1992, Cortes et Vapnik, 1995] sont probablement les classifieurs les plus répandus en apprentissage supervisé, communément présentés dans le cadre de la classification binaire avec $Y = \{-1, +1\}$. Signalons, qu'ils peuvent être étendus à différentes tâches d'apprentissage [12] comme la classification multiclasse [13] ou la régression. Tout au long de ce mémoire, les SVM seront utilisés comme modèle de référence lors des expérimentations menées.

Ils trouvent leur source dans la théorie de Vapnik-Chervonenkis et reposent sur deux notions clés : celle de marge maximale et celle de fonction noyau. Intuitivement, l'objectif est de trouver un séparateur linéaire dont la distance minimale — la marge — aux exemples d'apprentissage est maximale. Cependant, les données sont rarement linéairement séparables dans l'espace originel de description $X \in \mathbb{R}^d$. La solution est alors recourir à des noyaux de Mercer qui permettent de définir un espace de projection (implicite et potentiellement de dimension infinie) dans lequel les données seront linéairement séparables ou "presque". Avant de présenter cet espace, nous énonçons le principe des SVM dans l'espace X d'origine.

10. Voir [Yang et Jin, 2006, Bellet *et al.*, 2013b] pour un état de l'art.
11. *Support Vector Machine* en anglais.
12. Voir par exemple l'ouvrage [Schölkopf et Smola, 2002] pour plus de détails.
13. Le lecteur peut se référer à [Guermeur, 2007] pour un état de l'art sur les SVM multiclasse.

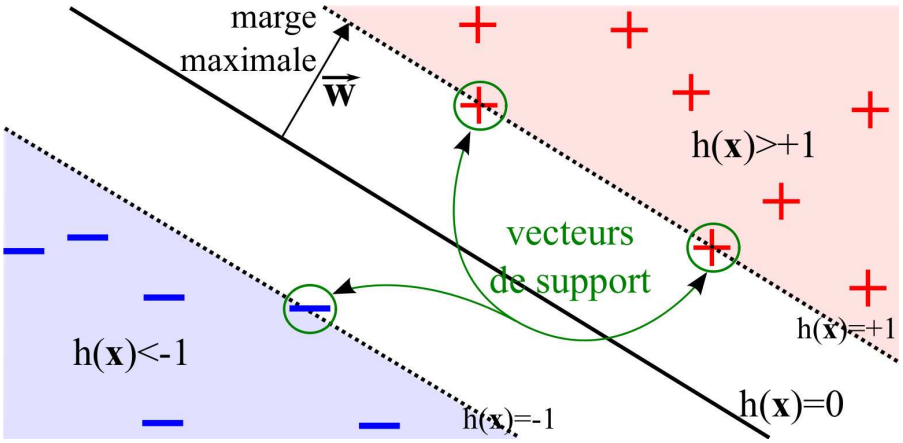


FIGURE 1.3 – *L'hyperplan optimal* $\mathbf{w}$ *séparant un échantillon avec une marge égale à* 1. *Les exemples encerclés sont les vecteurs de support de l'hyperplan* $\mathbf{w}$.

Dans l'espace de description X, la distance d'un exemple $\mathbf{x}'$ à l'hyperplan d'équation $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = 0$ dépend donc de $\langle \mathbf{w}, \mathbf{x}' \rangle$, car $\mathbf{w}$ est le vecteur directeur de l'hyperplan (il est donc orthogonal à $h(\mathbf{x})$). L'hyperplan optimal est en fait celui qui maximise la marge $\frac{\langle \mathbf{w}, \mathbf{x}' \rangle + b}{\|\mathbf{w}\|}$ (normalisée pour que tous les hyperplans soient comparables). La figure 1.3 illustre ce principe. Plus formellement, étant donné un échantillon d'apprentissage $S = (\mathbf{x}_i, y_i)_{i=1}^m$ la recherche du classifieur linéaire idéal s'exprime à l'aide du problème d'optimisation suivant en utilisant la forme canonique de l'hyperplan [14] :

$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{s.c.} & \forall i \in \{1, \ldots, m\},\ y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1, \end{cases} \tag{1.7}$$

où $\mathbf{w}$ est le vecteur directeur de l'hyperplan et b est un terme de biais. La fonction de classification est alors :

$$\text{sign} \left(\langle \mathbf{w}, \mathbf{x} \rangle + b \right).$$

La minimisation du problème (1.7) revient à trouver l'hyperplan de norme minimale — donc de plus grande marge — classant correctement la totalité des exemples d'apprentissage. Il existe donc une solution uniquement lorsque les données d'apprentissage sont linéairement séparables.

Lorsque les données ne sont pas linéairement séparables, la contrainte de non violation de la marge peut être relaxée et modélisée avec la fonction de perte hinge. Le problème d'optimisation prend la forme RRM (section 1.2.3) et devient :

$$\begin{cases} \min\limits_{\mathbf{w},b} & C \sum\limits_{i=1}^{m} \ell_{hinge}(h, (\mathbf{x}_i, y_i)) + \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{avec} & \forall i \in \{1, \ldots, m\},\ \ell_{hinge}(h, (\mathbf{x}_i, y_i)) = \left[1 - y_i \left(\langle \mathbf{w}, \mathbf{x} \rangle + b \right) \right]_+, \end{cases} \tag{1.8}$$

où $C > 0$ est le paramètre qui contrôle le compromis entre l'épaisseur de la marge et le nombre de violations de marge que l'on s'autorise. D'un point de vue pratique, la résolution des deux problèmes précédents est parfois inenvisageable car les calculs

14. La forme canonique d'un hyperplan revient à normaliser $\mathbf{w}$ et b de telle sorte que la marge soit égale à 1, c'est à dire : $\forall i \in \{1, \ldots, m\},\ y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1$.

sont de complexité $O(d^3)$, avec d la dimension de X. Pour simplifier la résolution, la théorie de l'optimisation stipule qu'un problème admet une forme duale équivalente lorsque la fonction objectif et les contraintes sont strictement convexes. Puisque c'est le cas ici, en dérivant le Lagrangien associé au problème (1.8), on obtient la forme duale suivante :

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} & \sum\limits_{i=1}^{m} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \\ \text{s.c.} & \forall i \in \{1, \ldots, m\},\ 0 \leq \alpha_i \leq C, \\ & \sum\limits_{i=1}^{m} y_i \alpha_i = 0. \end{cases} \tag{1.9}$$

Les multiplicateurs α_i sont appelés multiplicateurs de Lagrange. Lorsqu'ils sont non nuls, cela signifie qu'ils ont une influence dans la décision prise par le classifieur et, finalement, ce sont les exemples $\mathbf{x}_i$ associés qui sont appelés les vecteurs de support.

Jusqu'ici, nous avons supposé que les exemples d'apprentissage étaient linéairement séparables ou "presque" dans l'espace de description X. Cependant, cette situation idyllique n'arrive que très rarement. Un point important à souligner est que l'hyperplan solution et le problème d'optimisation font uniquement intervenir des produits scalaires $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ entre des vecteurs de X. Cela permet d'utiliser l'astuce offerte par les méthodes à noyaux appelée l'astuce du noyau [15]. En particulier, nous considérons les noyaux de Mercer :

Définition 1.10 (Noyau de Mercer) *Un fonction de similarité $K : X \times X \mapsto \mathbb{R}$ est un noyau de Mercer si elle est symétrique et semi-définie positive (SDP). Autrement dit, si elle vérifie :*

(i) $\forall (\mathbf{x}_i, \mathbf{x}_j) \in X^2,\ K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$,

(ii) pour tout entier positif fini $N \in \mathbb{R}^+$, pour tout $\alpha_1 \in \mathbb{R}$, ..., $\alpha_N \in \mathbb{R}$ et pour tout $\mathbf{x}_1 \in X$, ... $\mathbf{x}_N \in X$:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0.$$

En fait, ce sont les hypothèses de symétrie et de semi-définie positivité des noyaux de Mercer qui vont permettre l'application de l'astuce du noyau. En effet, la semi-définie positivité implique l'existence d'un espace de Hilbert implicite de dimension potentiellement infinie et auquel nous n'avons pas nécessairement accès. L'idée est donc de projeter les données dans cet espace pour qu'elles y soient linéairement séparables ou "presque". Cette projection s'effectue à l'aide d'une fonction non linéaire [16]. Plus formellement :

Théorème 1.5 (Théorème de Mercer) *Soit $X \subseteq \mathbb{R}^d$ un espace d'entrée. Un noyau de Mercer $K : X \times X \mapsto \mathbb{R}$ symétrique est semi-défini positif si et seulement s'il existe une fonction*

15. *Kernel trick* en anglais

16. Notons que la théorie des RKHS, que nous ne développons pas ici, permet, elle aussi, de "construire" cette fonction de plongement. Elle a, en outre, été étendue aux noyaux à valeurs opérateurs [Senkene et Tempel'man, 1973, Micchelli et Pontil, 2005]

de projection $\phi : X \mapsto \mathcal{H}_\phi$ (où $\mathcal{H}_\phi$ est un espace de Hilbert) telle que $K(\cdot,\cdot)$ puisse s'exprimer comme un produit scalaire :

$$K(\mathbf{x},\mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle .$$

Notons que la fonction $\phi(\cdot)$ est généralement inconnue et sous-jacente.

Concrètement, calculer le produit scalaire entre deux exemples $\mathbf{x}$ et $\mathbf{x}'$ dans l'espace de caractéristiques défini par $\mathcal{H}_\phi$ revient simplement à évaluer $K(\mathbf{x},\mathbf{x}')$ la valeur du noyau sur ces exemples. Cela évite le calcul explicite du plongement, et, finalement, étant donnés l'ensemble d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ et un noyau de Mercer $K(\cdot,\cdot)$, le problème d'optimisation des SVM devient :

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} & \sum\limits_{i=1}^m \alpha_i - \dfrac{1}{2}\sum\limits_{i=1}^m \sum\limits_{j=1}^m \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j), \\ \text{s.c.} & \forall i \in \{1,\dots,m\},\ 0 \leq \alpha_i \leq C, \\ & \sum\limits_{i=1}^m y_i \alpha_i = 0. \end{cases} \tag{1.10}$$

Dans cette situation, la règle de décision finale prend la forme :

$$\operatorname{sign}\left(\sum_{i=1}^m \alpha_i \langle \phi(\mathbf{x}), \phi(\mathbf{x}_i)\rangle + b\right) = \operatorname{sign}\left(\sum_{i=1}^m \alpha_i K(\mathbf{x},\mathbf{x}_i) + b\right).$$

Par la suite, un classifieur appris avec l'algorithme des SVM est appelé un classifieur-SVM.

Dans la littérature, il existe différents noyaux de Mercer classiques dont :

- Le noyau linéaire qui est la fonction noyau la plus simple. Il correspond au produit scalaire additionné à une constante c :

$$K(\mathbf{x}_i,\mathbf{x}_j) = {\mathbf{x}_i}^\top \mathbf{x}_j + c. \tag{1.11}$$

- Le noyau polynomial de degré $p \in \mathbb{N}^+$ utile lorsque les données sont normalisées :

$$K(\mathbf{x}_i,\mathbf{x}_j) = \left(a{\mathbf{x}_i}^\top \mathbf{x}_j + c\right)^p,$$

  Les paramètres a, c et p sont réglables.

- Le noyau gaussien est une fonction à base radiale d'épaisseur réglable σ, dont l'espace de projection est de dimension infinie :

$$K(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right). \tag{1.12}$$

Cependant, l'utilisation, le choix ou la construction/le réglage des noyaux de Mercer sont parfois complexes. D'une part, l'espace de projection est implicite, ce qui empêche d'y travailler directement. D'autre part, le choix du noyau dépend de la tâche considérée. Nous devons être capable de le régler correctement, voir même de le construire en amont, ce qui s'avère parfois difficile dû aux contraintes de symétrie et de semi-définie positivité. Dans la littérature, différentes approches existent pour contourner ces contraintes. Nous présentons dans la section suivante un de ces travaux.

1.4.3 Apprendre avec des fonctions de similarité (ϵ,γ,τ)-bonnes

Nous présentons maintenant le cadre théorique proposé par [Balcan *et al.*, 2008a, Balcan *et al.*, 2008b] permettant de contourner les contraintes liées à la symétrie et la semi-définie positivité des noyaux classiques dans le contexte de la classification binaire avec $Y = \{-1, +1\}$. Ce cadre introduit la notion de fonctions de similarité (ϵ,γ,τ)-bonnes suivante plus intuitive, plus flexible et plus facilement interprétable.

Définition 1.11 ([Balcan *et al.*, 2008a]) *Une fonction de similarité sur $X \in \mathbb{R}^d$ est une fonction $K : X \times X \to [-1, +1]$. $K(\cdot,\cdot)$ est dite (ϵ,γ,τ)-bonne sur un domaine P sur $X \times Y$, si il existe une fonction indicatrice aléatoire $R(\cdot)$ définissant un ensemble de points raisonnables tels que :*

(i) un taux de $1-\epsilon$ des exemples $(\mathbf{x}, y)$ vérifient :

$$\mathop{\mathbf{E}}_{(\mathbf{x}',y')\sim P} \left[yy' K(\mathbf{x},\mathbf{x}') | R(\mathbf{x}') = 1 \right] \geq \gamma,$$

(ii) $\mathop{\mathbf{Pr}}_{\mathbf{x}'\sim D} \left[R(\mathbf{x}') = 1 \right] \geq \tau.$

En d'autres termes :

(i) avec une marge γ, la majorité des exemples sont plus similaires aux points raisonnables de même classe qu'à ceux de classe opposée ;

(ii) au moins une proportion τ des points sont raisonnables.

Cette définition inclut donc aussi bien les noyaux de Mercer que des similarités ni semi-définies positives ni symétriques : les fonctions de similarité (ϵ,γ,τ)-bonnes sont donc plus générales que les noyaux de Mercer.
En pratique, les points raisonnables sont inconnus, il faut donc pouvoir les estimer à partir d'un ensemble aléatoire de r points dits *landmarks* noté $R = \{\mathbf{x}'_j\}_{j=1}^r$. Si $K(\cdot,\cdot)$ est (ϵ,γ,τ)-bonne sur un domaine P, alors *(i)* et *(ii)* sont des conditions suffisantes pour apprendre avec une grande probabilité un classifieur linéaire performant dans un espace appelé ϕ^R-espace et défini par la fonction de projection $\phi^R(\cdot)$ qui projette un point dans l'espace explicite de ses similarités à chacun des *landmarks* :

$$\phi^R : \left\{ \begin{array}{ccc} X & \to & \mathbb{R}^r \\ \mathbf{x} & \mapsto & (K(\mathbf{x},\mathbf{x}'_1), \ldots, K(\mathbf{x},\mathbf{x}'_r))^\top. \end{array} \right. \tag{1.13}$$

L'existence dans le ϕ^R-espace d'un tel classifieur linéaire est justifiée par le théorème suivant.

Théorème 1.6 ([Balcan *et al.*, 2008a]) *Soit $K(\cdot,\cdot)$ une fonction de similarité (ϵ,γ,τ)-bonne sur un domaine P sur $X \times Y$. Soit $R = \{\mathbf{x}'_j\}_{j=1}^r$ un échantillon de landmarks tirés i.i.d. selon D la distribution marginale de X tel que $r = \frac{2}{\tau}\left(\log(\frac{2}{\delta}) + 8\frac{\log(\frac{2}{\delta})}{\gamma^2}\right)$. Considérons la fonction de projection $\phi^R(\cdot)$ définie par l'équation* (1.13)*. Alors, pour tout $\delta \in (0, 1]$, avec une probabilité d'au moins $1-\delta$ sur le choix aléatoire de $R \sim (D)^r$, il existe un classifieur linéaire au plus $\epsilon + \delta$ au regard d'une marge d'au moins $\frac{\gamma}{2}$ dans l'espace induit par $\phi^R(\cdot)$.*

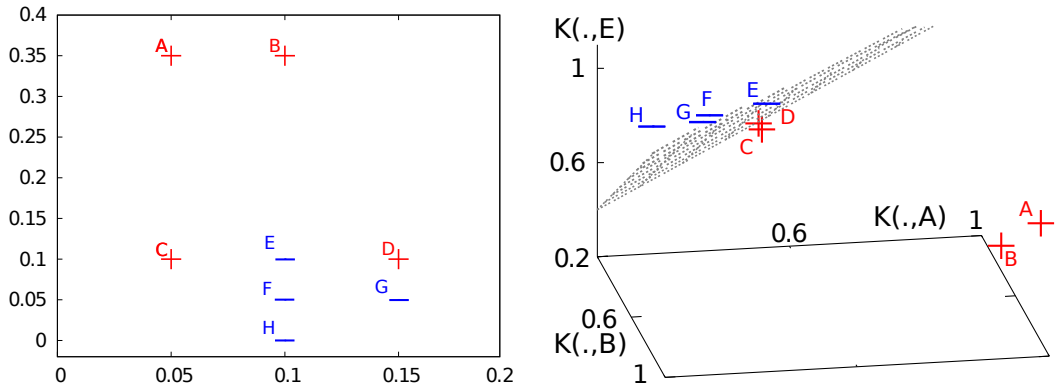


FIGURE 1.4 – *Exemple d'un ensemble de points positifs et négatifs. A gauche dans l'espace d'origine, à droite dans le ϕ^R-espace avec $R = \{A, B, E\}$, en gris l'hyperplan séparateur.*

	A	B	C	D	E	F	G	H
$K(\cdot, A)$	1	0.90	0.50	0.46	0.49	0.39	0.37	0.29
$K(\cdot, B)$	0.90	1	0.49	0.49	0.50	0.40	0.39	0.30
$K(\cdot, E)$	0.68	0.70	0.90	0.90	1	0.90	0.86	0.80
$Qualité((\mathbf{x}, y)) = \mathbb{E}_{R(\mathbf{x}')=1} yy' K(\mathbf{x}, \mathbf{x}')$	0.410	0.400	0.030	0.017	0.003	0.037	0.033	0.070

TABLE 1.1 – *Exemple des garanties d'une fonction de similarité : la distance à chaque point raisonnable et une évaluation des (ϵ,γ,τ)-garanties.*

Ainsi, étant donnés une fonction de similarité (ϵ,γ,τ)-bonne pour un problème de classification supervisée et — suffisamment — de *landmarks*, il existe avec une grande probabilité un classifieur linéaire d'erreur faible dans le ϕ^R-espace explicite.

Nous illustrons la notion de fonction de similarité (ϵ,γ,τ)-bonne à l'aide du petit exemple jouet suivant. Considérons un problème avec huit exemples étiquetés dans $[0,1] \times [0,1]$ soit unreprésentés sur la figure 1.4 :

- quatre positifs : $A = \big((0.05, 0.35), +1\big)$, $B = \big((0.1, 0.35), +1\big)$, $C = \big((0.05, 0.1), +1\big)$, $D = \big((0.15, 0.1), +1\big)$,
- quatre négatifs : $E = \big((0.1, 0.1), -1\big)$, $F = \big((0.1, 0.05), -1\big)$, $G = \big((0.15, 0.05), -1\big)$, $H = \big((0.1, 0), -1\big)$.

On peut remarquer que la position du point E empêche l'existence d'un classifieur linéaire d'erreur nulle. Soit la fonction de similarité :

$$K(\mathbf{x}, \mathbf{x}') = 1 - 2\|\mathbf{x} - \mathbf{x}'\|_2,$$

où $\|\mathbf{x} - \mathbf{x}'\|_2$ est la distance euclidienne classique. Nous supposons que le domaine P est uniforme sur les huits exemples, alors en admettant que trois des huit exemples sont raisonnables (nous verrons par la suite comment estimer les points raisonnables), A, B et E, τ correspond à $\frac{3}{8}$. Les (ϵ,γ,τ)-garanties de K peuvent alors être évaluées à l'aide de la définition 1.11. Les valeurs obtenues sont reportées dans la table 1.1.

- Si la marge γ égale 0.002, nous remarquons que la marge associée à chacun des exemples est supérieure à γ ce qui rend la similarité $(0, 0.002, 3/8)$-bonne.
- Cependant, si l'on pose $\gamma = 0.02$, la similarité est alors $(0.25, 0.02, 3/8)$-puisque deux exemples n'atteignent pas une marge supérieure à 0.02.

Finalement dans l'espace de projection défini par les similarités aux trois points raisonnables, $\phi^R(\cdot) = (K(\cdot, A), K(\cdot, B), K(\cdot, E))^\top$ il existe un classifieur parfait : $\text{sign}\,[K(\cdot, A) + K(\cdot, B) - K(\cdot, C)]$. Ce classifieur n'est probablement pas le meilleur, le but ici est de proposer une illustration simple des propriétés de la définition 1.11.

Le critère donné par la définition 1.11 requiert la minimisation du nombre de violations de la marge dont l'approximation est en pratique NP-difficile. De manière similaire aux SVM, les auteurs ont proposé de considérer une relaxation de la définition 1.11 basée sur la fonction de perte hinge.

Définition 1.12 ([Balcan *et al.*, 2008a]) *Une fonction de similarité $K(\cdot, \cdot)$ est (ϵ, γ, τ)-bonne au sens de la fonction de perte hinge sur un domaine P sur $X \times Y$, s'il existe une fonction indicatrice (aléatoire) $R(\cdot)$ définissant un ensemble de points raisonnables tels que :*

(i) on a :

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \ell_{hinge}(h(\mathbf{x}), y) \leq \epsilon,$$

où $h(\mathbf{x}) = \dfrac{1}{\gamma} \mathop{\mathbf{E}}_{(\mathbf{x}',y')\sim P} \left[y' K(\mathbf{x}, \mathbf{x}') | R(\mathbf{x}') = 1\right]$ *et $\ell_{hinge}(\cdot, \cdot)$ est la fonction de perte hinge,*

(ii) $\mathop{\mathbf{Pr}}_{\mathbf{x}'\sim D} \left[R(\mathbf{x}')\right] \geq \tau.$

Toujours en utilisant le ϕ^R-espace explicite défini dans l'équation (1.13), les auteurs ont prouvé :

Théorème 1.7 ([Balcan *et al.*, 2008a]) *Soit $K(\cdot, \cdot)$ une (ϵ, γ, τ)-bonne fonction de similarité au sens de la perte hinge sur un domaine P sur $X \times Y$. Pour tout $\epsilon_1 > 0$ et $0 < \delta < \frac{\gamma\epsilon_1}{4}$, soit $R = \{\mathbf{x}'_j\}_{j=1}^r$ un échantillon de $r = \frac{2}{\tau}\left(\log(2/\delta) + 16\frac{\log(2/\delta)}{(\gamma\epsilon_1)^2}\right)$ landmarks tirés i.i.d. selon D. En considérant la projection $\phi^R(\cdot)$ définie par l'équation* (1.13)*, avec une probabilité d'au moins $1 - \delta$ sur le choix aléatoire de $R \sim (D)^r$, il existe un classifieur linéaire de risque au plus $\epsilon + \epsilon_1$ au sens de la fonction de perte hinge, avec une marge de γ dans l'espace induit par $\phi^R(\cdot)$.*

Finalement, étant donnés S un échantillon de m points étiquetés tirés *i.i.d.* selon le domaine P, et r *landmarks*, un séparateur linéaire $\boldsymbol{\alpha} \in \mathbb{R}^r$ peut être trouvé efficacement en résolvant un programme linéaire où la fonction objectif cherche à minimiser la quantité de violation de la marge relachée par la fonction de perte hinge. Nous en donnons ici une formulation de type RRM (section 1.2.3) équivalente [17] à celle proposée dans

17. Le problème d'optimisation originel donné dans [Balcan *et al.*, 2008a] s'exprime avec une contrainte en norme 1. Pour simplifier sa résolution, nous utilisons ici une version équivalente exprimée à l'aide d'une régularisation en norme 1.

[Balcan *et al.*, 2008a].

$$\begin{cases} \min_{\boldsymbol{\alpha}} \frac{1}{m} \sum_{i=1}^{m} \ell_{hinge}(h(\mathbf{x}_i), y_i)) + \lambda \|\boldsymbol{\alpha}\|_1, \\ \text{avec } \ell_{hinge}(h(\mathbf{x}_i), y_i) = \left[1 - y_i \sum_{j=1}^{r} \alpha_j K(\mathbf{x}_i, \mathbf{x}'_j) \right]_+ , \end{cases} \tag{1.14}$$

où λ est le paramètre de compromis. Par la suite, un classifieur linéaire appris en résolvant le problème 1.14 est appelé un classifieur-SF [18]. Soulignons que ce problème est très proche du problème d'optimisation pour les SVM, à trois différences près.

- La fonction $K(\cdot, \cdot)$ n'a ni besoin d'être symétrique, ni SDP.
- L'hyperplan séparateur vit dans un plongement explicite construit à partir des similarités $K(\cdot, \mathbf{x}')$, contrairement aux SVM pour lesquels l'espace de plongement est un espace de Hilbert, en pratique implicite, induit par un noyau de Mercer.
- La régularisation du problème est de type $\| \cdot \|_1$ sur $\boldsymbol{\alpha}$, qui est une contrainte de parcimonie sur le modèle inféré. Elle permet d'approximer les points raisonnables à partir des *landmarks* (ceux choisis sont associés à un poids $\alpha_j \neq 0$).

Dans le chapitre 6, ce sont ces trois propriétés qui vont nous permettre de définir un algorithme pour construire un espace explicite où l'objectif sera de "rapprocher" deux domaines différents tout en gardant de bonnes garanties en généralisation.

1.5 Synthèse

Dans ce chapitre, nous avons introduit les concepts de l'apprentissage supervisé, le vocabulaire et les notations que nous utiliserons dans la suite de ce manuscrit. De plus, nous avons décrit trois algorithmes : les k-PPV, les SVM et l'apprentissage de classifieurs-SF. Dans ce cadre classique d'apprentissage supervisé, les données d'apprentissage sont supposées représentatives des données à étiqueter. Cependant, cette hypothèse est parfois difficile à vérifier. Dans le chapitre suivant, nous énonçons le cadre de l'adaptation de domaine, qui nous intéressera tout particulièrement dans la partie III, pour lequel la distribution des exemples d'apprentissage n'est pas la même que la distribution des nouvelles données de test sur lesquelles doit s'appliquer le modèle.

18. SF pour *Similarity Function*.

Adaptation de domaine 2

Dans le chapitre précédent, nous nous sommes placés dans l'un des cadres théoriques les plus communs en apprentissage automatique : l'apprentissage s'effectue à partir d'un échantillon représentatif des données que l'on désire classer. Bien que cette hypothèse soit parfois une bonne approximation de la réalité, elle reste difficile à vérifier en pratique. Reprenons l'exemple du système de filtrage de spams évoqué en introduction : un système performant pour un utilisateur donné ne le sera pas nécessairement pour un utilisateur recevant des e-mails de natures différentes. Il faut alors être capable d'adapter le système d'un utilisateur à un autre. On peut modéliser ce problème d'un point de vue statistique : l'échantillon d'apprentissage n'est alors plus issu de la même distribution de probabilité que les données que l'on désire traiter, on parle d'adaptation de domaine [1]. L'objectif est alors de proposer des méthodes pour adapter un modèle d'un ou de plusieurs domaine(s) source vers un domaine cible différent. De nos jours, cette problématique s'avère

1. *Domain adaptation* en anglais, voir [Jiang, 2008, Margolis, 2011] pour un état de l'art.

très active en apprentissage automatique, mais aussi dans de nombreuses communautés scientifiques telles qu'en multimédia (*ex :* [Duan *et al.*, 2009, Roy *et al.*, 2012]), en traitement d'images (*ex :* [Saenko *et al.*, 2010]), en traitement automatique de la langue (*ex :* [Daumé III, 2007, Daumé III *et al.*, 2010, McClosky *et al.*, 2006]), en bioinformatique (*ex :* [Liu *et al.*, 2008]). En effet, entre la grande diversité des données accessibles par Internet et le fait que la personnalisation soit au cœur de beaucoup de problématiques, toutes ces communautés s'intéressent à tirer au mieux parti de toutes les informations disponibles afin d'adapter ou de transférer les connaissances dont on dispose sur de nouveaux types de données.

L'abondance des travaux en adaptation de domaine implique qu'il est difficile d'en faire un état de l'art détaillé et complet [2]. Dans ce chapitre, nous n'avons donc pas la prétention de proposer un état de l'art exhaustif de l'adaptation de domaine. Nous faisons le choix de nous focaliser sur les travaux qui en représentent pour nous les contributions majeures. Nous présentons les grands principes en section 2.1, les travaux théoriques fondateurs, sur lesquels nos contributions de la partie III se basent, en section 2.2, puis deux algorithmes, en section 2.3, que nous utiliserons dans nos comparaisons expérimentales.

2.1 Qu'est ce que l'adaptation de domaine ?

2.1.1 Un des champs d'étude de l'apprentissage par transfert

La problématique portée par l'adaptation de domaine fait partie d'une thématique bien plus vaste appelée l'apprentissage par transfert [3]. L'apprentissage par transfert peut être vu comme la capacité d'un système à reconnaître et appliquer des connaissances et des compétences, apprises à partir de tâches antérieures, sur de nouvelles tâches ou domaines partageant des similitudes. La question qui se pose est : comment identifier les similitudes entre la ou les tâche(s) cible(s) et la ou les tâche(s) source(s), puis comment transférer la connaissance de la ou des tâche(s) source(s) vers la tâche(s) cible(s) ? En fonction des hypothèses supposées sur les tâches et les domaines, on distingue différents types d'apprentissage par transfert résumés sur la figure 2.1 et énoncés ci-dessous.

- Lorsque les domaines sont identiques mais que les tâches sont différentes, c'est-à-dire lorsque les ensembles d'étiquetages sont différents, on parle d'apprentissage par transfert inductif.
- Lorsque les domaines et les tâches sont différents, on parle d'apprentissage par transfert non supervisé.

2. Le lecteur peut se référer à différents tutoriaux proposés dans différents domaines, tels que par exemple : `http://adaptationtutorial.blitzer.com/` (ICML 2010), `http://vc.sce.ntu.edu.sg/transferlearning.html` (CVPR 2012), `http://interspeech2012.org/DomainAdaptation.html` (InterSpeech 2012).

3. *Transfer learning* en anglais. Le lecteur peut se référer à [Pan et Yang, 2010] pour un état de l'art sur l'apprentissage par transfert.

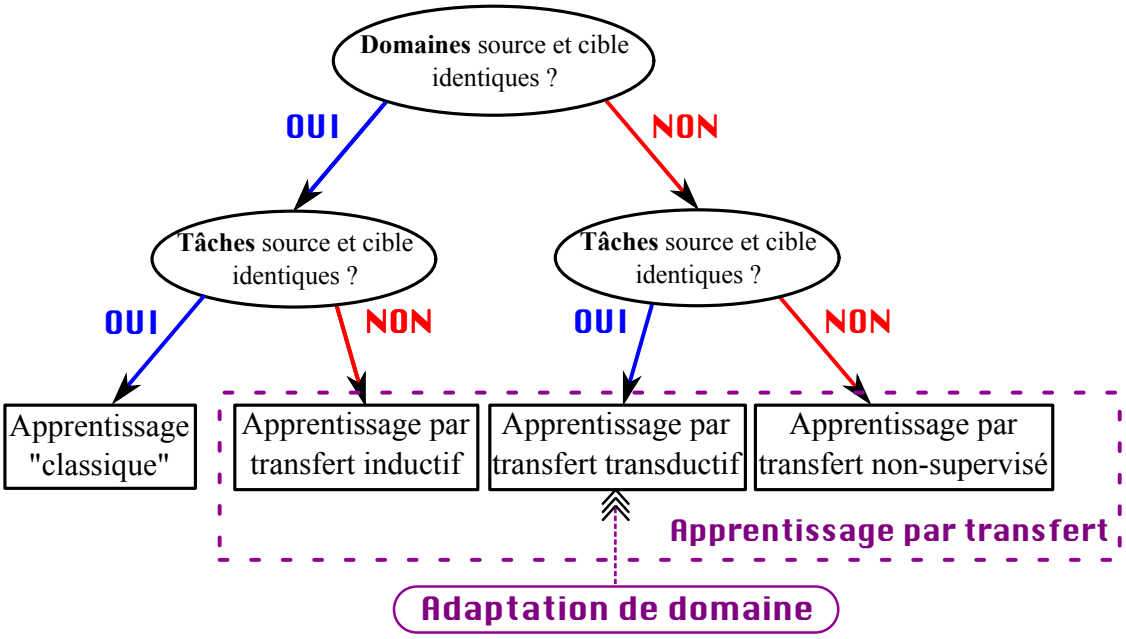


FIGURE 2.1 – *Distinction entre l'apprentissage classique et l'apprentissage par transfert, et positionnement de l'adaptation de domaine.*

- Lorsque les domaines sont différents mais que les tâches sont identiques, on parle d'apprentissage par transfert transductif. C'est ici que se place l'adaptation de domaine.

Dans ce manuscrit, nous nous focalisons sur l'adaptation de domaine où l'objectif est d'effectuer une tâche d'adaptation d'un domaine source vers un domaine cible. Notons que lorsque plusieurs domaines sources sont disponibles, on parle d'adaptation de domaine multi-source [Crammer *et al.*, 2008].

Dans ce qui suit, nous présentons plus en détail l'adaptation de domaine.

2.1.2 Un peu de formalisme

Nous allons distinguer deux contextes de l'adaptation de domaine qui diffèrent par l'information dont on dispose sur le domaine cible.

- Adaptation de domaine non supervisée : l'ensemble d'apprentissage est constitué d'un ensemble de données étiquetées sources, d'un ensemble de données sources non étiquetées et d'un ensemble de données cibles non étiquetées. Signalons, en pratique, que l'ensemble de données non étiquetées sources peut simplement correspondre à l'ensemble étiqueté privé de ses étiquettes.

- Adaptation de domaine semi-supervisée : en complément des informations accessibles en adaptation de domaine non supervisée, on dispose d'un ensemble — de petite taille — de données étiquetées cibles[4]. Permettre l'utilisation d'un

4. Signalons que certains auteurs, comme [Daumé III, 2007], parlent également d'adaptation de domaine supervisée lorsque toutes les données cibles disponibles sont étiquetées, mais ce contexte relativement peu fréquent, est assez peu étudié dans la littérature.

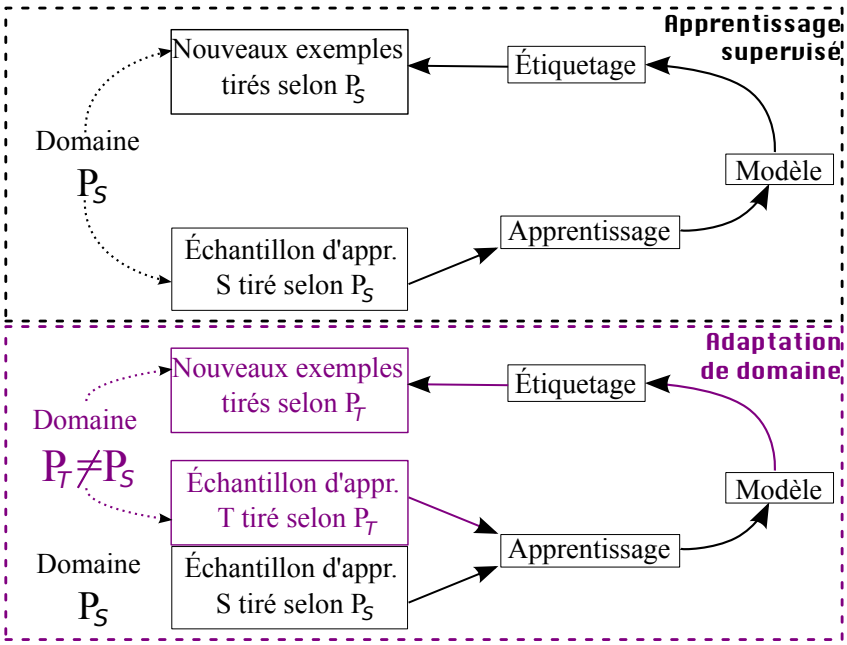


FIGURE 2.2 – *Distinction entre l'apprentissage supervisé et l'adaptation de domaine.*

tel ensemble va faciliter l'adaptation, puisque l'on prend en compte une connaissance informative sur le domaine cible. Cette tâche n'a évidemment de sens que si l'ensemble étiqueté cible est trop peu informatif pour pouvoir inférer, à lui seul, une hypothèse performante.

Plus formellement, la principale différence entre l'apprentissage supervisé classique et l'adaptation de domaine, illustrée par la figure 2.2, réside dans l'étude de deux domaines P_S et P_T différents (fixes et inconnus) sur $X \times Y$. D_S et D_T sont les distributions marginales sur X respectives. La tâche d'adaptation consiste alors à transférer nos connaissances du domaine source P_S vers le domaine cible P_T.

- Pour ce faire, en adaptation de domaine non supervisée, nous considérons un échantillon d'apprentissage source $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m^s}$ de m^s exemples tirés *i.i.d.* selon P_S. En outre, nous disposons d'un échantillon source non-étiqueté $S_u = \{\mathbf{x}_i^s\}_{i=1}^{m_u^s}$ de m_u^s exemples tirés *i.i.d.* selon D_S, et d'un échantillon cible non étiqueté $T_u = \{\mathbf{x}_i^t\}_{i=1}^{m_u^t}$ de m_u^t exemples tirés *i.i.d.* selon D_T. En pratique, S_u peut correspondre à S privé de ses étiquettes : $S_u = \{\mathbf{x}_i^s | (\mathbf{x}_i^s, y_i^s) \in S\}$; et les échantillons peuvent être de taille égale : $m_u^t = m_u^s = m_u$.

- En adaptation de domaine semi supervisée, nous permettons l'utilisation d'un échantillon étiqueté cible $T = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{m^t}$ consistué de m^t exemples cibles tirés *i.i.d.* selon P_T et tel que : $m^t \ll m^s$.

Étant donné un ensemble d'hypothèses $\mathcal{H}$ de X vers Y, l'objectif de l'adaptation de domaine est de trouver l'hypothèse issue de $\mathcal{H}$ qui minimise le risque réel sur le domaine cible $\mathbf{R}_{P_T}(\cdot)$, ou au moins d'en trouver une bonne approximation. La question majeure soulevée par ce problème est la suivante : si un modèle a été appris sur un domaine source, quelle sera sa capacité en généralisation sur le domaine cible ? Avant

de présenter les garanties classiques en adaptation de domaine en section 2.2, nous énonçons les trois grands principes algorithmiques qui se distinguent, puis quelques situations d'adaptation particulières.

2.1.3 Les grands types d'algorithmes

Les algorithmes de repondération

L'approche la plus intuitive est celle de la repondération en fonction des données, dont le but est, communément, de repondérer l'échantillon étiqueté source de tel sorte qu'il "ressemble" le plus possible à l'échantillon cible au sens de la fonction de perte considérée [Huang *et al.*, 2007, Jiang et Zhai, 2007, Mansour *et al.*, 2009b]. Concrètement, la repondération repose sur l'affectation de poids à la fonction de perte, relativement aux exemples, avec pour objectif la minimisation du risque réel sur le domaine cible. Selon le principe ERM en apprentissage supervisé classique, énoncé dans la section 1.2.1 du chapitre précédent, une solution pour définir une repondération adéquate à l'adaptation de domaine est la suivante. Étant donnés une fonction de perte $\ell(\cdot,\cdot)$ et un espace d'hypothèses $\mathcal{H}$, l'objectif est de trouver le minimiseur du risque sur le domaine cible. En supposant que $\text{supp}(P_T) \subseteq \text{supp}(P_S)$, où $\text{supp}(P)$ est le support de P, alors pour toutes les hypothèses h issues de $\mathcal{H}$, le risque réel $\mathbf{R}^{\ell}_{P_T}(h)$ se ré-écrit :

$$\begin{aligned}
\mathbf{R}^{\ell}_{P_T}(h) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_T} \ell\left(h(\mathbf{x}),y\right) \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_T} \frac{P_S(\mathbf{x},y)}{P_S(\mathbf{x},y)} \ell\left(h(\mathbf{x}),y\right) \\
&= \sum_{(\mathbf{x},y)\in(X\times Y)} P_T(\mathbf{x},y) \frac{P_S(\mathbf{x},y)}{P_S(\mathbf{x},y)} \ell\left(h(\mathbf{x}),y\right) \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \frac{P_T(\mathbf{x},y)}{P_S(\mathbf{x},y)} \ell\left(h(\mathbf{x}),y\right). \qquad (2.1)
\end{aligned}$$

Le poids, à affecter à la fonction de perte appliquée sur un $(\mathbf{x},y)$, est donc défini comme le ratio $\frac{P_T(\mathbf{x},y)}{P_S(\mathbf{x},y)}$. Il implique ainsi une solution naturelle à l'adaptation de domaine. Néanmoins, il est impossible de calculer la valeur exacte de ce ratio pour un exemple $(\mathbf{x},y)$, en particulier lorsque la quantité d'étiquettes cibles n'est pas suffisante. Dans de nombreux travaux, comme dans [Lin *et al.*, 2002, Kubat *et al.*, 1997], il est alors souvent supposé que les domaines partagent la même distribution conditionnelle selon Y, c'est-à-dire que : $P_S(\mathbf{x}|y) = P_T(\mathbf{x}|y)$. De la même manière, il est possible de supposer que ce sont les distributions des classes qui sont les mêmes, c'est-à-dire que : $P_S(y|\mathbf{x}) = P_T(y|\mathbf{x})$. Cette dernière situation, qui a fait l'objet de nombreux travaux, est connue sous le nom de *covariate-shift* ou de biais de sélection d'échantillon[5] [Shimodaira, 2000, Zadrozny, 2004, Huang *et al.*, 2007]. Nous détaillons ce cadre particulier en section 2.1.4.

5. *sample selection biais* en anglais.

Les algorithmes d'auto-étiquetage

Un second principe algorithmique pour l'adaptation de domaine est l'auto-étiquetage itératif des données cibles. Le principe est simple :

- un modèle h est appris uniquement à l'aide des données étiquetées disponibles ;
- h est utilisé pour étiqueter les données cibles $\mathbf{x}^t$ par $h(\mathbf{x}^t)$, ceci correspond à l'auto-étiquetage ;
- un nouveau modèle est appris à partir de ces données auto-étiquetées. En général, on n'en traite qu'un sous-ensemble et le processus est ré-itéré jusqu'à ce que l'auto-étiquetage apparaisse le plus juste possible.

Nous pouvons, par exemple, citer l'algorithme DASVM [Bruzzone et Marconcini, 2010] que nous présenterons en détail en section 2.3.1. À chaque itération, étant donné le classifieur-SVM appris à l'itération précédente, DASVM modifie l'ensemble étiqueté : on insère des données cibles auto-étiquetées et on supprime les données sources les plus éloignées de l'hyperplan. Un nouveau classifieur-SVM est appris à partir de cet échantillon modifié.

Notons que dans la littérature, il existe des méthodes itératives ne nécessitant pas d'auto-étiquetage. Elles sont, par exemple, basées sur le principe du *boosting*, mais requièrent généralement des données cibles étiquetées [Dai *et al.*, 2007, Yao et Doretto, 2010, Habrard *et al.*, 2013].

Les algorithmes de recherche d'un espace de représentation commun

La dernière approche que nous présentons est celle qui se rapproche le plus de nos contributions de la partie III : la recherche ou la construction d'un espace de représentation commun aux deux domaines. Elle trouve principalement sa justification théorique dans les bornes en généralisation d'adaptation de domaine que nous détaillons dans la section 2.2. L'objectif est d'inférer un espace de représentation dans lequel les domaines seront similaires tout en gardant de bonnes performances sur le domaine source.

Un des premiers algorithmes suivant ce principe est l'algorithme SCL (*Structural Correspondance Learning* [Blitzer *et al.*, 2006]) qui cherche une représentation de rang faible en tirant parti des données non étiquetées issues des deux domaines. Pour ce faire, on identifie dans cet espace des attributs fréquents dans les deux domaines et montrant un comportement similaire. Ils sont alors appelés les attributs pivots et vont permettent de mettre en correspondance les domaines. Finalement, l'hypothèse est apprise dans l'espace originel augmenté de la représentation définie par ces correspondances.

En gardant la même intuition, [Blitzer *et al.*, 2011] ont proposé l'algorithme *Coupled Subspace*. Alors que SCL cherche un simple espace de projection en faisant correspondre des attributs communs aux domaines, *Coupled Subspace* apprend deux projections, une pour chaque domaine, afin de coupler les attributs partagés avec ceux spécifiques aux

domaines. Deux hypothèses sont alors apprises, une sur le sous-espace spécifique au domaine source et une sur celui spécifique au domaine cible, puis elles sont fusionnées pour former l'hypothèse finale.
Une autre solution vise à la sélection d'un sous-ensemble des attributs partagés par les domaines. C'est ce que réalise l'algorithme CODA (*CO-training dor Domain Adaptation*) [Chen *et al.*, 2011a] que nous décrivons en section 2.3.2.
D'autres approches se basent sur une augmentation de l'espace de description. C'est le cas de l'algorithme *EasyAdapt++* [Daumé III, 2007, Daumé III *et al.*, 2010] apprenant simultanément une hypothèse générale, une hypothèse source et une hypothèse cible dans un espace augmenté. Cet espace est défini par l'ensemble des attributs de l'espace d'origine et est obtenu en concaténant une représentation commune, une représentation source et une représentation cible. Plus formellement, les données étiquetées sources et cibles sont respectivement plongées par les fonctions :

$$\phi_S(\mathbf{x}) = (\mathbf{x}, \mathbf{x}, \mathbf{0})^\top, \qquad \phi_T(\mathbf{x}) = (\mathbf{x}, \mathbf{0}, \mathbf{x})^\top.$$

Si l'on dispose de données non étiquetées, on fait appel à la fonction :

$$\phi_U(\mathbf{x}) = (\mathbf{0}, \mathbf{x}, -\mathbf{x})^\top$$

Finalement, un classifieur linéaire est appris dans cet espace augmenté.

2.1.4 Quelques situations particulières

Quelle que soit l'approche choisie, l'adaptation de domaine reste un problème difficile. Elle est connue pour être une tâche assez complexe à résoudre lorsqu'aucune hypothèse n'est supposée sur la tâche considérée [Ben-David *et al.*, 2010, Ben-David et Urner, 2012]. Cependant, nous pouvons distinguer des contextes pouvant — parfois — faciliter l'adaptation.

covariate shift

L'hypothèse du *covariate shift* [Shimodaira, 2000] est une des hypothèses les plus fréquemment utilisées en adaptation de domaine[6]. Concrètement, nous rappelons qu'elle est vérifiée lorsque les deux domaines partagent la même fonction d'étiquetage. Autrement dit, étant donnée une observation $\mathbf{x}$, les distributions conditionnelles source et cible de Y selon X sont les mêmes, mais les distributions marginales sur X peuvent être différentes. Plus formellement, on suppose :

$$\begin{aligned} \forall \mathbf{x} \in X,\ & P_S(y|\mathbf{x}) = P_T(y|\mathbf{x}), \\ & \text{mais :}\ D_S(\mathbf{x}) \neq D_T(\mathbf{x}). \end{aligned}$$

Alors que cette hypothèse semble très forte, elle fait sens pour de nombreuses tâches d'adaptation. Considérons, par exemple, une tâche d'étiquetage morpho-syntaxique[7]

6. Le *covariate shift* est parfois considéré comme une problématique à part entière différente de l'adaptation de domaine.

7. *Part of speech tagging* en anglais.

en traitement automatique de la langue. Si on applique un modèle, appris à partir d'un domaine (par exemple des articles scientifiques), sur un nouveau domaine (par exemple des articles de journaux), il est raisonnable de supposer que la différence entre les deux tâches réside uniquement dans les distributions marginales sur les mots de la langue française plutôt que sur l'étiquetage de chaque mot.

Sous l'hypothèse du *covariate shift*, l'adaptation de domaine apparaît donc simple à résoudre. En effet, il facilite le problème de la repondération présenté en section 2.1.3. Lorsque $P_T(y|\mathbf{x}) = P_S(y|\mathbf{x})$, le ratio $\frac{P_T(\mathbf{x},y)}{P_S(\mathbf{x},y)}$, obtenu dans l'équation (2.1), peut être réécrit de la manière suivante :

$$\begin{aligned}\frac{P_T(\mathbf{x},y)}{P_S(\mathbf{x},y)} &= \frac{D_T(\mathbf{x})P_T(y|\mathbf{x})}{D_S(\mathbf{x})P_S(y|\mathbf{x})} \\ &= \frac{D_T(\mathbf{x})}{D_S(\mathbf{x})}.\end{aligned}$$

Dans ce cas, il suffit donc de pondérer la fonction de perte associée à un exemple par le ratio $\frac{D_T(\mathbf{x})}{D_S(\mathbf{x})}$.

Cependant, même sous cette hypothèse, le problème d'adaptation peut rester difficile à résoudre, en particulier lorsque la fonction d'étiquetage commune aux deux domaines est en dehors de la portée de l'échantillon étiqueté source [Ben-David *et al.*, 2010]. Nous en verrons une illustration en section 2.2.5.

Ratio entre distributions marginales

Comme suggéré ci-dessus, lorsque les domaines source et cible vivent dans des régions totalement disjointes, l'adaptation peut rapidement devenir difficile. Afin de se prévenir d'un tel scénario, on peut supposer qu'il existe une minoration non nulle sur le ratio de la densité en chaque point[8]. Cependant, cette hypothèse est difficile à vérifier dans la réalité, en particulier lorsque des données sont spécifiques à un domaine et n'apparaissent jamais dans l'autre. En pratique on relâche cette hypothèse et on suppose vérifiée la définition suivante :

Définition 2.1 ([Ben-David *et al.*, 2012b]) *Soit $\mathcal{B} \subseteq 2^X$ une collection de sous-ensembles de l'espace X. Soit $\epsilon > 0$, on définit le ϵ-ratio entre les distributions marginales sur X source et cible par rapport à $\mathcal{B}$ par :*

$$W_{\mathcal{B},\epsilon}(D_S, D_T) = \inf_{\substack{\mathbf{b}\in\mathcal{B} \\ D_T(\mathbf{b})\geq\epsilon}} \frac{D_S(\mathbf{b})}{D_T(\mathbf{b})}.$$

De plus, on définit le ratio entre la marginale source et la marginale cible par rapport à $\mathcal{B}$ par :

$$W_{\mathcal{B}}(D_S, D_T) = \inf_{\substack{\mathbf{b}\in\mathcal{B} \\ D_T(\mathbf{b})\neq 0}} \frac{D_S(\mathbf{b})}{D_T(\mathbf{b})}.$$

Cette mesure devient pertinente pour l'adaptation de domaine lorsque le ratio peut être minoré loin de 0.

8. *Pointwise density ratio* en anglais.

On peut remarquer que ce ratio peut être obtenu en posant : $\mathcal{B} = \{\{\mathbf{x}\} : \mathbf{x} \in X\}$. Pour tout $\mathcal{B} \subseteq 2^X$, on a : $W_{\{\{\mathbf{x}\}:\mathbf{x}\in X\}}(D_S, D_T) \leq W_{\mathcal{B}}(D_S, D_T)$, ainsi minorer le ratio en chaque point, loin de 0, est la plus forte des restrictions.

Probabilité lipschitzienne

Un autre contexte permet de relâcher les deux hypothèses précédentes : celui de la probabilité lipschitzienne. La probabilité lipschitzienne généralise la notion standard des fonctions λ-lipschitziennes définies pour des fonctions à valeurs réelles $f(\cdot)$ par : $\forall (x, x') \in \mathbb{R}^2,\ |f(x) - f(x')| \leq \lambda \|x - x'\|$, pour une constante λ. Cette condition peut aisément être appliquée à un étiquetage probabiliste. Cependant, si la fonction d'étiquetage est déterministe, cela implique que deux points d'étiquettes différentes doivent être à une distance d'au moins $\frac{1}{\lambda}$. Cette restriction assez forte peut être relachée grâce à la définition suivante.

Définition 2.2 (Probabilité lipschitzienne [Urner *et al.*, 2011, Ben-David *et al.*, 2012b]) *Soit une fonction croissante $\psi : \mathbb{R}^+ \mapsto [0, 1]$. On dit que $f : X \mapsto \{-1, +1\}$ est ψ-Lipschitzienne par rapport à une distribution D sur X si pour tout $\lambda > 0$, on a :*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim D}\left(\exists \mathbf{x}' \ :\ f(\mathbf{x}) \neq f(\mathbf{x}') \wedge \left\|\mathbf{x} - \mathbf{x}'\right\| \leq \lambda\right) \leq \psi(\lambda).$$

Notons que si une fonction est λ-Lipschitizienne alors elle est ψ-Lipschitzienne par rapport à toute distribution en posant :

$$\psi(x) = \begin{cases} 0 & \text{si } x \leq \lambda \\ 1 & \text{si } x > \lambda. \end{cases}$$

Si $f(\cdot)$ correspond à la vraie fonction d'étiquetage du domaine P, alors cette définition peut, en quelque sorte, formaliser l'hypothèse selon laquelle la frontière de décision vit dans une région de densité faible [9], très commune en apprentissage semi-supervisé. Autrement dit, les données sont séparables en différents clusters étiquetés de manière homogène.

Sous ces deux dernières hypothèses (probabilité lipschitzienne et ratio entre distributions marginales), il est possible de dériver une borne en généralisation sur l'erreur du 1-PPV [Urner *et al.*, 2011, Ben-David *et al.*, 2012b].

λ-shift

[Mansour et Schain, 2012] ont récemment proposé la notion de *λ-shift* qui utilise les propriétés des algorithmes robustes [Xu et Mannor, 2010, Xu et Mannor, 2012] présentées en section 1.3.4 du chapitre 1. Pour rappel, la notion de robustesse algorithmique en apprentissage supervisé offre des garanties de convergence lorsque dans chaque région $X_i \times Y_j$ la fonction de perte de l'algorithme renvoie des valeurs similaires à un $\epsilon > 0$ près. Dans le cas de l'adaptation de domaine, l'idée est de considérer

9. *cluster assumption* en anglais.

les distributions conditionnelles selon l'étiquetage dans une région X_i. On va, encore une fois, supposer que des points proches sont de même étiquette, et ce avec un certain biais. Le *λ-shift* est formellement défini par :

Définition 2.3 (*λ-shift* [Mansour et Schain, 2012]) *Soit P_S et P_T les domaines source et cible sur $X \times Y$. On dit que $P_T(y|\mathbf{x})$ et $P_S(y|\mathbf{x})$ sont liées par l'hypothèse du λ-shift, noté $P_T(y|\mathbf{x}) \in \lambda\,(P_S(y|\mathbf{x}))$ si pour tout $y \in Y$, on a :*

$$P_T(y|\mathbf{x}) \leq P_S(y|\mathbf{x}) + \lambda\,(1 - P_S(y|\mathbf{x}))\,,$$
$$\text{et } P_T(y|\mathbf{x}) \geq P_S(y|\mathbf{x})(1-\lambda).$$

Le *λ-shift* restreint donc le changement de la probabilité cible d'une étiquette : ce changement est d'au plus une proportion λ de la probabilité source des autres étiquettes ou une proportion $1-\lambda$ de la probabilité source de l'étiquette. En ce sens, cette définition peut être vue comme une généralisation du *covariate-shift*, pour lequel on suppose l'égalité entre les distributions conditionnelles.

Sous cette hypothèse, [Mansour et Schain, 2012] ont dérivé une borne en généralisation pour l'adaptation de domaine et en ont déduit des variantes adaptatives des SVM en classification binaire et en régression.

Dans la partie III de cette thèse, nous nous attaquons au problème de l'adaptation de domaine dans toute sa généralité, sans supposer de telles hypothèses. Nous présentons donc dans la section suivante, les résultats théoriques permettant de dériver des bornes en généralisation pour l'adaptation de domaine.

2.2 Garanties en généralisation pour l'adaptation de domaine

Nous exposons maintenant les travaux fondateurs de la théorie de l'adaptation de domaine en apprentissage automatique.

2.2.1 Nécessité d'une mesure de divergence entre les domaines

Nous rappelons que le but de l'adaptation de domaine est de trouver une hypothèse admettant une erreur cible faible même lorsqu'aucune information sur les étiquettes cibles n'est disponible. Comme nous le verrons dans la section 2.2.5, ce problème peut clairement se révéler difficile à résoudre même sous des hypothèses fortes [Ben-David et Urner, 2012, Ben-David *et al.*, 2010]. Pour dériver des bornes en généralisation pour l'adaptation de domaine, il est essentiel de capturer la différence entre les domaines souce et cible : plus les domaines sont similaires, plus l'adaptation est aisée (voir figure 2.3).

Concrètement, les deux domaines P_S et P_T diffèrent si leurs marginales D_S et D_T sur X sont différentes, ou si la fonction d'étiquetage source diffère de la fonction cible, ou si les deux cas précédents se produisent. Ceci suggère de prendre en compte deux

divergences : une entre les marginales D_S et D_T et l'autre entre les étiquetages. En adaptation de domaine semi-supervisée, lorsque des étiquettes cibles sont disponibles, les deux mesures peuvent être combinées (comme dans [C. Zhang, 2012]). Dans le cas contraire, il est préférable de les différencier puisque dans une telle situation il sera impossible d'estimer le meilleur étiquetage cible. Habituellement, comme suggéré dans la section précédente, on admet qu'il existe un lien entre l'étiquetage source et l'étiquetage cible. Une solution consiste alors à chercher une représentation pour laquelle D_S et D_T sont proches tout en gardant de bonnes garanties sur le domaine source. Autrement dit, en supposant qu'il existe aussi un lien entre les marginales, nous allons chercher à le reconstruire.

Dans la littérature, différentes mesures existent pour estimer à quel point deux distributions diffèrent. Nous pouvons citer, la $\mathcal{A}$-divergence [Kifer *et al.*, 2004] proposée à l'origine pour mesurer des changements dans des flux de données, la *perturbed variation* [Harel et Mannor, 2012] qui permet de donner une information sur la similarité de deux distributions en calculant une mesure de perturbation, ou encore la divergence de Kullback-Leibler [Kullback et Leibler, 1951] qui est un cas particulier des divergences de Bregman [Bregman, 1967] et qui sont des mesures de distorsions indéxées par fonctions strictement convexes. Notons que la divergence de Kullback-Leibler est au centre de la théorie PAC-Bayésienne détaillée dans le chapitre suivant.

De plus, dans le cadre spécifique de l'adaptation de domaine, où l'on veut calculer la distance entre deux domaines différents, nous pouvons citer en exemple les travaux suivants. Signalons que ces mesures peuvent toutes être utilisées lorsque l'on dispose de plusieurs domaines sources.

- La $\mathcal{H}$-divergence proposée dans [Ben-David *et al.*, 2007, Ben-David *et al.*, 2010] s'inspire de la $\mathcal{A}$-divergence et constitue le fondement de la théorie de l'adaptation de domaine en classification binaire avec la fonction de perte $0-1$. Cette divergence a été généralisée à la régression et à des fonctions de pertes plus générales par [Mansour *et al.*, 2009a]. Nous la présentons plus en détail dans la section suivante.

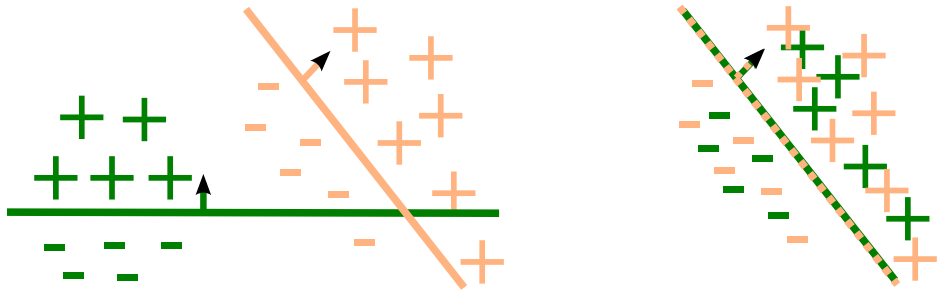
- Lorsque plusieurs domaines sources sont accessibles, nous pouvons spécifiquement faire appel à la divergence de Rényi [Mansour *et al.*, 2009b].

- [C. Zhang, 2012] ont proposé une mesure de divergence uniquement valable en adaptation de domaine semi-supervisée.

- La *divergence prior* Bayésienne [Li et Bilmes, 2007] qui, dans un contexte Bayésien, se focalise sur les hypothèses proches de l'hypothèse source optimale.

- Dans le cas particulier du *covariate-shift*, on peut faire appel aux travaux liés au *two-sample test* comme par exemple dans [Huang *et al.*, 2007].

Notons que dans le chapitre 7, nous proposerons une mesure de divergence appropriée à l'apprentissage de votes de majorité.



(a) *Les domaines sont très différents : un classifieur appris sur le domaine source ne sera pas performant sur le domaine cible.*

(b) *Les domaines sont similaires : le classifieur appris depuis le domaine source est performant sur les deux domaines.*

FIGURE 2.3 – *Illustration de la nécessité de mesurer la similarité entre domaines. Le domaine source est en vert foncé (pos. +, neg. −), le domaine cible en orange clair.*

Nous énonçons maintenant les deux premiers travaux fondateurs de la théorie de l'adaptation de domaine en apprentissage automatique qui se basent sur une divergence entre marginales et qui constituent la base de nos contributions développées dans la partie III.

2.2.2 Une divergence entre les distributions marginales pour la classification binaire

Tout d'abord, [Ben-David *et al.*, 2007, Ben-David *et al.*, 2010] ont proposé, dans le cadre de la classification binaire en considérant la fonction de perte $0-1$, une mesure de divergence appelée la $\mathcal{H}$-divergence et basée sur la définition de la $\mathcal{A}$-divergence [Kifer *et al.*, 2004]. [Mansour *et al.*, 2009a] ont ensuite étendu cette définition à la *discrepancy* permettant de considérer des fonctions de pertes plus générales. Ces deux divergences, équivalentes lorsque l'on parle de la fonction de perte $0-1$, calculent la valeur maximale, pour deux classifieurs issus de $\mathcal{H}$, de la différence entre les désaccords (selon la fonction de perte) source et cible (voir la définition 1.3). Nous en énonçons les deux définitions.

Définition 2.4 (la $\mathcal{H}$-divergence [Ben-David *et al.*, 2007, Ben-David *et al.*, 2010] (version simplifiée)) *Étant donnée une classe d'hypothèses $\mathcal{H}$. Soit D_S et D_T deux distributions sur X, alors la $\mathcal{H}$-divergence entre D_S et D_T est définie par :*

$$\frac{1}{2}d_{\mathcal{H}}(D_S, D_T) = \sup_{(h,h')\in\mathcal{H}^2} \left|\mathbf{R}_{D_T}(h,h') - \mathbf{R}_{D_S}(h,h')\right|,$$

où $\mathbf{R}_D(h,h') = \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{I}\left(h(\mathbf{x}) \neq h'(\mathbf{x})\right)$ est le désaccord entre h et h' sur la distribution D.

Dans sa forme originelle, cette définition est uniquement valable pour la fonction de perte $0-1$. Cependant, en l'énonçant comme ci-dessus, il apparaît simple de la généraliser à d'autres fonctions de perte :

Définition 2.5 (la *discrepancy* [Mansour *et al.*, 2009a]) *Étant donnée une classe d'hypothèses $\mathcal{H}$. Soit $\ell : Y \times Y \mapsto \mathbb{R}^+$ une fonction de perte. Soit D_S et D_T deux distributions sur X, alors la*

discrepancy entre D_S et D_T est définie par :

$$\mathrm{disc}_\ell(D_S, D_T) = \sup_{(h,h')\in\mathcal{H}^2} \left| \mathbf{R}^\ell_{D_T}(h,h') - \mathbf{R}^\ell_{D_S}(h,h') \right|,$$

où $\mathbf{R}^\ell_D(h,h') = \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \ell(h(\mathbf{x}), h'(\mathbf{x}))$ est le désaccord entre h et h' relativement à la fonction de perte $\ell(\cdot,\cdot)$.

Ces deux définitions sont donc clairement équivalentes lorsque la fonction de perte est la fonction de perte $0-1$:

$$\mathrm{disc}_{\ell_{0\text{-}1}}(D_S, D_T) = \tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T).$$

Néanmoins, les analyses de l'adaptation de domaine qui en résultent sont différentes. Lorsque l'on parle de la $\mathcal{H}$-divergence, [Ben-David *et al.*, 2007, Ben-David *et al.*, 2010] ont étudié la consistance du processus de minimisation empirique de la divergence en se basant sur la VC-dim. (section 1.3.1, chapitre 1), alors que pour la *discrepancy*, [Mansour *et al.*, 2009a] se sont basés sur la complexité de Rademacher (section 1.3.2, chapitre 1). Les deux analyses correspondent respectivement aux théorèmes suivants.

Théorème 2.1 ([Ben-David *et al.*, 2007, Ben-David *et al.*, 2010]) *Soit $\mathcal{H}$ un espace d'hypothèses de VC-dim. $\mathrm{VC}(\mathcal{H})$ finie. Si S_u et T_u sont des échantillons de taille $m^t_u = m^s_u = m_u$ dont les éléments sont respectivement tirés i.i.d selon les distribution D_S et D_T sur X et si $\frac{1}{2} d_{\mathcal{H}}(S_u, T_u)$ est la $\mathcal{H}$-divergence empirique mesurée sur les échantillons S_u et T_u, alors pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u \sim (D_S)^{m_u}$ et $T_u \sim (D_T)^{m_u}$, on a :*

$$\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) \leq \tfrac{1}{2} d_{\mathcal{H}}(S_u, T_u) + 4\sqrt{\frac{2\,\mathrm{VC}(\mathcal{H})\log(2m_u) + \log\frac{2}{\delta}}{m_u}}.$$

L'analyse en complexité de Rademacher aboutit, quant à elle, au résultat suivant.

Théorème 2.2 ([Mansour *et al.*, 2009a]) *Soit $\mathcal{H}$ un espace d'hypothèse. Si S_u et T_u sont des échantillons de taille $m^t_u = m^s_u = m_u$ dont les éléments sont respectivement tirés i.i.d selon les distribution D_S et D_T et si $\mathrm{disc}_{\ell_{0\text{-}1}}(S_u, T_u)$ est la discrepancy empirique associée à la fonction de perte $0-1$ et mesurée sur les échantillons S_u et T_u, alors pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u \sim (D_S)^{m_u}$ et $T_u \sim (D_T)^{m_u}$, on a :*

$$\mathrm{disc}_{\ell_{0\text{-}1}}(D_S, D_T) \leq \mathrm{disc}_{\ell_{0\text{-}1}}(S_u, T_u) + 4\left(\mathfrak{R}_S(\mathcal{H}) + \mathfrak{R}_T(\mathcal{H})\right) + 6\sqrt{\frac{\log\frac{4}{\delta}}{2m_u}},$$

où $\mathfrak{R}_S(\mathcal{H})$, respectivement $\mathfrak{R}_T(\mathcal{H})$, est la complexité empirique de Rademacher sur S, respectivement sur T.

Ces résultats montrent que la *discrepancy* empirique associée à la fonction de perte $0-1$, c'est-à-dire la $\mathcal{H}$-divergence empirique, converge uniformément vers sa vraie valeur. Notons qu'un résultat similaire existe dans le cas de fonctions de perte associées au problème de régression [10]. Les deux théorèmes précédents justifient donc de la consistance du processus de minimisation empirique de la divergence.

10. Nous invitons le lecteur à se reporter à [Mansour *et al.*, 2009a, Cortes et Mohri, 2011] pour plus de détails.

Notons qu'en pratique, la $\mathcal{H}$-divergence empirique peut s'estimer en cherchant un classifieur qui vise à séparer la marginale source de la marginale cible. Pour ce faire, on étiquette les exemples sources en -1 et les exemples cibles en $+1$, puis on cherche à apprendre un classifieur discriminant les instances sources des cibles. La $\mathcal{H}$-divergence empirique se calcule alors directement à partir de l'erreur de ce classifieur.

Lemme 2.1 ([Ben-David *et al.*, 2007, Ben-David *et al.*, 2010]) *Pour tout espace d'hypothèses $\mathcal{H}$ de X vers Y et deux échantillons non étiquetés S_u et T_u de taille $m_u^t = m_u^s = m_u$, on a :*

$$\frac{1}{2} d_{\mathcal{H}}(S_u, T_u) = 1 - \min_{h \in \mathcal{H}} \left[\frac{1}{m_u} \sum_{\mathbf{x}: h(\mathbf{x})=-1} \mathbf{I}\left[\mathbf{x} \in S_u\right] + \frac{1}{m_u} \sum_{\mathbf{x}: h(\mathbf{x})=+1} \mathbf{I}\left[\mathbf{x} \in T_u\right] \right].$$

L'inconvénient, ici, est que l'évaluation se réalise à partir du classifieur minimisant cette erreur, ce qui est en général NP-dur. Dans la littérature, les auteurs approximent généralement cette distance via l'apprentissage d'un classifieur linéaire avec une perte légèrement modifée ou par un classifieur-SVM.

2.2.3 Bornes en généralisation pour l'adaptation de domaine

Les résultats précédents permettent d'obtenir des garanties en généralisation lorsque l'on cherche à apprendre un modèle à partir du domaine source. C'est, par exemple, le cas de l'approche basique ERM, présentée en section 1.2.1, sous condition que $\mathcal{H}$ ait une complexité finie (soit en terme de VC-dim., soit de complexité de Rademacher). Nous présentons maintenant ces garanties.

Tout d'abord, [Ben-David *et al.*, 2007, Ben-David *et al.*, 2010] ont prouvé la borne d'adaptation suivante qui s'avère précise lorsqu'il existe un classifieur dans $\mathcal{H}$ à la fois performant sur le domaine source et le domaine cible.

Théorème 2.3 ([Ben-David *et al.*, 2007, Ben-David *et al.*, 2010]) *Soit P_S et P_T deux domaines sur $X \times Y$ dont D_S et D_T sont les distributions marginales respectives sur X. Soit $\mathcal{H}$ une classe d'hypothèses, alors on a :*

$$\forall h \in \mathcal{H},\ \mathbf{R}_{P_T}(h) \ \leq\ \mathbf{R}_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu, \tag{2.2}$$

où $\nu = \mathbf{R}_{P_S}(h^*) + \mathbf{R}_{P_T}(h^*)$ l'erreur jointe optimale, avec $h^* = \operatorname*{argmin}_{h \in \mathcal{H}} \left(\mathbf{R}_{P_S}(h) + \mathbf{R}_{P_T}(h)\right)$ l'hypothèse jointe optimale.

Cette borne dépend de trois termes. $\mathbf{R}_{P_S}(h)$ est l'erreur classique, en apprentissage supervisé, mesurée sur le domaine souce. La divergence $\frac{1}{2} d_{\mathcal{H}}(D_S, D_T)$ dépend de la classe d'hypothèses $\mathcal{H}$ et correspond à l'écart maximal entre les désaccords source et cible entre deux hypothèses de $\mathcal{H}$. Autrement dit, elle quantifie à quel point les hypothèses de $\mathcal{H}$ peuvent "détecter" les différences entre les marginales D_S et D_T, sans information sur l'étiquetage. Le dernier terme ν est relié à la meilleure hypothèse h^* sur les deux domaines en même temps et mesure la qualité de $\mathcal{H}$ en fonction de l'étiquetage. Si h^* est mauvaise, il sera complexe trouver une hypothèse performante

sur le domaine cible. Enfin, l'équation (2.2) combinée avec la théorie VC exprime un compromis entre la performance source d'une hypothèse h de $\mathcal{H}$, la complexité de $\mathcal{H}$ et "l'incapacité" des hypothèses de $\mathcal{H}$ à détecter les différences entre les domaines. Plus formellement, on a le théorème suivant.

Théorème 2.4 ([Ben-David *et al.*, 2007, Ben-David *et al.*, 2010]) *Soit P_S et P_T deux domaines sur $X \times Y$ dont D_S et D_T sont les marginales respectives sur X. Soit $\mathcal{H}$ un espace d'hypothèses de VC-dim. finie $VC(\mathcal{H})$. Soit S_u et T_u des échantillons de taille m_u dont les éléments sont respectivement tirés i.i.d. selon les distributions D_S et D_T. Soit S un échantillon étiqueté source composé de m^s exemples sources tirés i.i.d selon P_S. Alors, pour tout $\delta \in (0, 1]$, avec une probabilité d'au moins $1 - \delta$ sur le choix des échantillons aléatoires $S \sim (P_S)^{m^s}$, $S_u \sim (D_S)^{m_u}$ et $T_u \sim (D_T)^{m_u}$, on a :*

$$\forall h \in \mathcal{H},\ \mathbf{R}_{P_T}(h) \leq \mathbf{R}_S(h) + \sqrt{\frac{\mathrm{VC}(\mathcal{H})\left(\ln \frac{2m^s}{\mathrm{VC}(\mathcal{H})} + 1\right) + \ln\left(\frac{4}{\delta}\right)}{m^s}} + \tfrac{1}{2} d_{\mathcal{H}}(S_u, T_u) + 4\sqrt{\frac{2\,\mathrm{VC}(\mathcal{H}) \ln(2m_u) + \ln\left(\frac{2}{\delta}\right)}{m_u}} + \nu.$$

Similairement, [Mansour *et al.*, 2009a] ont proposé la borne d'adaptation de domaine suivante basée sur la *discrepancy* pour toute fonction de perte $\ell(\cdot, \cdot)$ symétrique et vérifiant l'inégalité triangulaire. Nous ne présentons ici que la formulation pour la fonction de perte $0 - 1$.

Théorème 2.5 (Corollaire du théorème 8 de [Mansour *et al.*, 2009a]) *Soit P_S et P_T deux domaines sur $X \times Y$ dont D_S et D_T sont les marginales respectives sur X. Soit $\mathcal{H}$ une classe d'hypothèses, alors on a :*

$$\forall h \in \mathcal{H},\ \mathbf{R}_{P_T}(h) \leq \mathbf{R}_{D_S}(h_S^*, h) + \mathrm{disc}_{\ell_{0-1}}(D_S, D_T) + v, \tag{2.3}$$

où $v = \mathbf{R}_{P_T}(h_T^*) + \mathbf{R}_{D_S}(h_S^*, h_T^*)$ et $h_T^* = \operatorname{argmin}_{h \in \mathcal{H}} \mathbf{R}_{P_T}(h)$ et $h_S^* = \operatorname{argmin}_{h \in \mathcal{H}} \mathbf{R}_{P_S}(h)$ sont respectivement les hypothèses optimales sur le domaine cible et le domaine source.

Dans ce contexte, la borne (2.3) majore directement [11] l'écart entre l'erreur cible d'un classifieur $\mathbf{R}_{P_T}(h)$ et celle de l'hypothèse cible optimale $\mathbf{R}_{P_T}(h_T^*)$. Combinée avec une analyse en terme de complexité de Rademacher, cette borne s'exprime, dans le théorème suivant, comme un compromis entre le désaccord entre l'hypothèse h considérée et la meilleure hypothèse source h_S^*, la complexité de Rademacher de $\mathcal{H}$ et — encore une fois — la divergence entre les deux domaines mesurée par $\mathrm{disc}_{\ell_{0-1}}(D_S, D_T)$.

Théorème 2.6 ([Mansour *et al.*, 2009a]) *Soit P_S et P_T deux domaines sur $X \times Y$ dont D_S et D_T sont les distributions marginales respectives sur X. Soit $\mathcal{H}$ un espace d'hypothèses. Soit S_u et T_u des échantillons de taille m_u dont les éléments sont respectivement tirés i.i.d. selon les distributions D_S et D_T. Soit S un échantillon étiqueté source composé de m^s exemples sources tirés i.i.d selon P_S. Alors, pour tout $\delta \in [0, 1)$, avec une probabilité d'au moins $1 - \delta$ sur le choix des*

11. Il suffit de passer le terme $\mathbf{R}_{P_T}(h_T^*)$ de l'autre côté de l'inégalité.

échantillons aléatoires $S \sim (P_S)^{m^s}$, $S_u \sim (D_S)^{m_u}$ *et* $T_u \sim (D_T)^{m_u}$, *on a :*

$$\forall h \in \mathcal{H},\ \mathbf{R}_{P_T}(h) \leq \mathbf{R}_S(h^*_S, h) + \mathrm{disc}_{\ell_{0\text{-}1}}(S_u, T_u) + \upsilon + 4\left(\Re_{S_u}(\mathcal{H}) + \Re_{T_u}(\mathcal{H})\right) + 6\sqrt{\frac{\log\frac{8}{\delta}}{2m_u}} + \Re_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{8}{\delta}}{2m^s}}. \quad (2.4)$$

Pour conclure, les deux approches énoncent deux bornes en généralisation pour l'adaptation de domaine, difficilement comparables, mais impliquant la même intuition : un algorithme d'adaptation de domaine doit être capable d'inférer un espace de représentation dans lequel les deux domaines tendent à être indiscernables tout en gardant de bonnes performances sur le domaine source. En pratique, chercher un tel espace en minimisant la divergence et l'erreur source en même temps est difficile. Nous nous attaquerons à cet inconvénient, dans le chapitre 7, en proposant une analyse originale de l'adaptation de domaine pour minimiser conjointement la divergence entre les domaines et l'erreur sur le domaine source dans le cadre des votes de majorité qui est au cœur de ce mémoire.

Nous présentons maintenant un moyen de prendre en considération quelques étiquettes cibles lors de la phase d'apprentissage.

2.2.4 Extension à l'adaptation de domaine semi-supervisée

Il arrive parfois que des étiquettes cibles soient disponibles. Afin de tirer bénéfice de cette information plus que pertinente, [Ben-David *et al.*, 2010] ont étendu leur analyse pour les considérer. Nous rappelons que l'échantillon étiqueté (S, T) est alors décomposé en deux sous-ensembles : $S = \{(\mathbf{x}^s_i, y^s_i)\}_{i=1}^{m^s}$ est l'échantillon source constitué de m^s exemples étiquetés *i.i.d.* selon P_S et $T = \{(\mathbf{x}^t_i, y^t_i)\}_{i=1}^{m^t}$ est l'échantillon cible constitué de m^t exemples étiquetés *i.i.d.* selon P_T. Soit $\theta \in [0, 1]$ tel que $m^t = \theta m$ et $m^s = (1-\theta)m$ impliquant que (S, T) soit composé de $m = m^s + m^t$ exemples étiquetés tels que $m^t \ll m^s$. La minimisation de l'erreur empirique cible $\mathbf{R}_T(\cdot)$ n'est alors pas la meilleure solution car T n'est pas représentatif du domaine cible P_T. Une solution envisager est alors de minimiser la combinaison convexe des erreurs empiriques source et cible définie par :

$$\mathbf{R}^{\kappa}_{(S,T)}(h) = \kappa\ \mathbf{R}_T(h) + (1-\kappa)\ \mathbf{R}_S(h), \quad (2.5)$$

où $\kappa \in [0, 1]$ contrôle le compromis erreur empirique cible et erreur empirique source. L'erreur réelle pondérée associée est : $\kappa\ \mathbf{R}_{P_T}(h) + (1-\kappa)\mathbf{R}_{P_S}(h)$.

Le théorème suivant énonce les garanties en généralisation dans une telle situation.

Théorème 2.7 ([Ben-David *et al.*, 2010]) *Soit* P_S *et* P_T *deux domaines sur* $X \times Y$ *dont* D_S *et* D_T *sont les distributions marginales respectives sur* X. *Soit* $\mathcal{H}$ *un espace d'hypothèses de VC-dim.* $\mathrm{VC}(\mathcal{H})$ *finie. Soit* S_u *et* T_u *des échantillons de taille* m_u *dont les éléments sont respectivement tirés i.i.d selon les distribution* D_S *et* D_T. *Soit* (S, T) *un échantillon étiqueté de taille* m *généré en tirant aléatoirement* θm *exemples depuis* P_T *et* $(1-\theta)m$ *depuis* P_S. *Si* h_S *est le minimiseur empirique de* $\mathbf{R}^{\kappa}_{(S,T)}(h)$ *sur* (S, T) *et* $h^*_T = \min_{h\in\mathcal{H}} \mathbf{R}_{P_T}(h)$ *l'erreur cible optimale, alors*

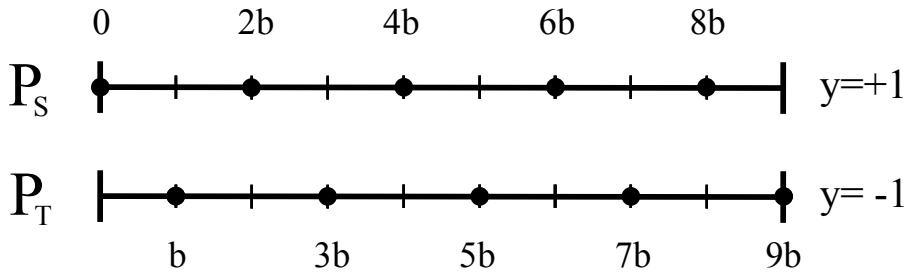


FIGURE 2.4 – *Représentation des domaines source et cible P_S et P_T dans une situation de covariate shift inadéquate pour l'adaptation de domaine. Pour un $b \in (0,1)$ fixé, le domaine source P_S est la distribution uniforme sur $\{2kb \ : \ k \in \mathbb{N},\ 2kb \leq 1\} \times \{+1\}$ et la distribution cible P_T est la distribution uniforme sur $\{(2k+1)b \ : \ k \in \mathbb{N},\ (2k+1)b \leq 1\} \times \{-1\}$.*

alors pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix des échantillons aléatoires, on a :

$$\mathbf{R}_{P_T}(h_S) \leq \mathbf{R}_{P_T}(h_T^*) + 4\sqrt{\frac{\kappa^2}{\theta} + \frac{(1-\kappa)^2}{1-\theta}}\sqrt{\frac{2\,\mathrm{VC}(\mathcal{H})\log\left(2(m+1)\right) + 2\log\frac{8}{\delta}}{m}} + 2(1-\theta)\left(\tfrac{1}{2}d_{\mathcal{H}}(S_u, T_u) + 4\sqrt{\frac{2\,\mathrm{VC}(\mathcal{H})\log(2m_u) + \log\frac{8}{\delta}}{m_u}} + \nu\right).$$

Remarquons que si $\kappa = 0$, on ignore les données cibles, et on retombe sur le théorème 2.3 avec une estimation empirique pour l'erreur sur le domaine source. Si $\kappa = 1$, alors ce sont les données sources que l'on ignore, la borne devient une borne de classification supervisée usuelle sur le domaine cible. Lorsque κ vaut sa valeur optimale, la borne est toujours au moins plus précise que l'une de ces deux situations. Finalement, on peut remarquer qu'en faisant varier κ, la borne permet de contrôler le compromis entre erreur source et erreur cible.

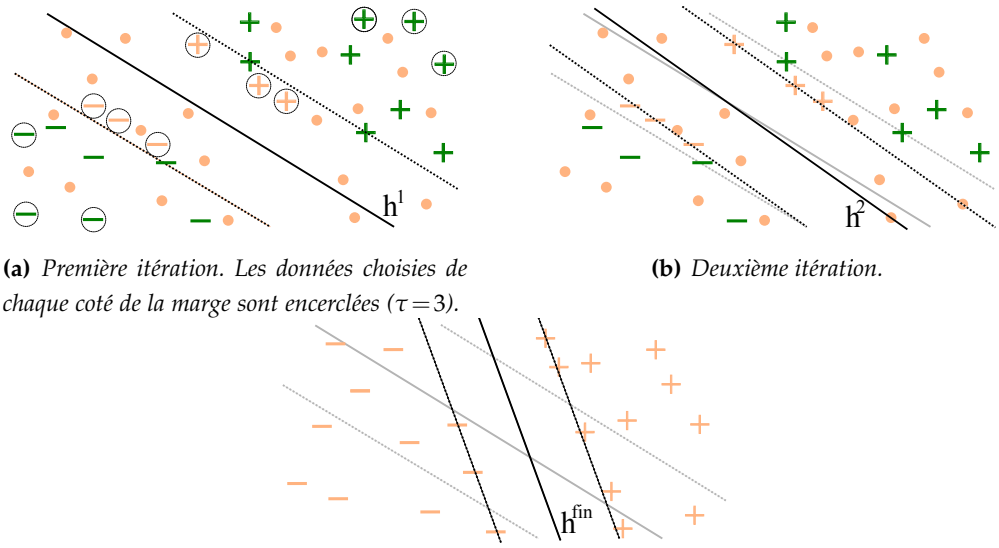
Ce cadre théorique permet donc de considérer des étiquettes cibles lors de la phase d'apprentissage. Nous en tirerons bénéfice en chapitre 6 pour étendre notre contribution à l'adaptation de domaine semi-supervisée.

2.2.5 Illustration de la difficulté de l'adaptation de domaine

Quelle que soit l'analyse, les bornes d'adaptation peuvent parfois être très imprécises et rien ne nous assure d'adapter correctement. Prenons l'exemple issu de [Ben-David *et al.*, 2010] en illustration (voir la figure 2.4). Soit $b \in (0,1)$, le domaine source P_T est la distribution uniforme sur $\{2kb \ : \ k \in \mathbb{N},\ 2kb \leq 1\} \times \{+1\}$ et la distribution cible P_T est la distribution uniforme sur $\{(2k+1)b \ : \ k \in \mathbb{N},\ (2k+1)b \leq 1\} \times \{-1\}$. On pose $\mathcal{H}$ comme l'espace d'hypothèses de fonctions de seuil. On définit pout tout $t \in [0,1]$ une fonction de seuil h_t par :

$$\forall x \in \mathbb{R},\ h_t(x) = \begin{cases} +1 & \text{si } x < t, \\ -1 & \text{sinon.} \end{cases}$$

Ainsi : $\mathcal{H} = \{h_t \ : \ t \in [0,1]\}$. Dans une telle situtation, on a donc :



(a) *Première itération. Les données choisies de chaque coté de la marge sont encerclées ($\tau = 3$).*

(b) *Deuxième itération.*

(c) *Dernière itération.*

FIGURE 2.5 – *L'hyperplan séparateur (ligne noire) et la marge (en pointillés noirs) à différentes itérations de DASVM sur un exemple jouet. Les données sources sont en vert foncé (+ pos., − neg.). Les données cibles sont en orange clair : les auto-étiquettes sont représentées par + et − et les données non étiquetées par un rond. Les lignes grises sur* **(b)** *et* **(c)** *rappellent l'hyperplan appris à la première itération.*

- l'hypothèse de *covariate shift* est vérifiable pour P_S et P_T ;
- la divergence vaut $\frac{1}{2} d_{\mathcal{H}}(P_S, P_T) = |9b - 8b| = b$ (b peut être arbitrairement faible) ;
- la meilleure hypothèse h_S^* sur le domaine source est d'erreur nulle, la meilleure h_T^* sur le domaine cible est aussi d'erreur nulle. Sur la figure 2.4, on a par exemple :

$$h_S^* = \begin{cases} +1 & \text{si } x \geq 0 \\ -1 & \text{sinon,} \end{cases} \quad \text{et} \quad h_T^* = \begin{cases} -1 & \text{si } x < 9b \\ +1 & \text{sinon ;} \end{cases}$$

- pour la borne (2.2), on a : $\nu = 1 - b$;
  pour la borne (2.3), on a : $\upsilon = 1$.

Les deux derniers points rendent les bornes d'adaptation imprécises : elles sont supérieures à 1. Ainsi, l'hypothèse de *covariate shift* ne garantit pas nécessairement une adaptation correcte. Au final, h_S^*, dont l'erreur sur le domaine cible vaut 1, est l'hypothèse qui minimise la borne.

2.3 Exemples d'algorithmes

Il existe de nombreux algorithmes d'adaptation de domaine. Dans cette section, nous présentons ceux auxquels nous ferons appel lors des expérimentations menées dans les chapitres 6 et 7, puis nous énonçons une méthode de validation des paramètres.

2.3.1 DASVM : un algorithme d'adaptation itératif

Algorithme 2 DASVM

entrée un échantillon étiqueté source S, un échantillon non-étiqueté cible T_u, hyperparamètres τ et β

sortie Un classifieur h_{DASVM}

$S^1 \leftarrow S$; $T_u^1 \leftarrow T_u$; $\mathit{it} \leftarrow 0$

repéter

$\mathit{it} \leftarrow \mathit{it} + 1$

$h^{\mathit{it}} \leftarrow$ apprentissage d'un classifieur-SVM à partir de S^{it}

$T_u^{\mathit{it}+1} \leftarrow \left\{ (\mathbf{x}_i^t, h^{\mathit{it}}(\mathbf{x}_i^t) \right\}_{i=1}^{|T_u^{\mathit{it}}|}$

$f^{\mathit{it}} \leftarrow$ fonction score associée à h^{it}

$H_{up}^{\mathit{it}} \leftarrow \left\{ (\mathbf{x}_i^t, h^{\mathit{it}}(\mathbf{x}_i^t)) \in T_u^{\mathit{it}+1} : 1 \geq f^{\mathit{it}}(\mathbf{x}_i^t) \geq 0 \right\}$ (par ordre décroissant selon la marge)

$H_{low}^{\mathit{it}} \leftarrow \left\{ (\mathbf{x}_i^t, h^{\mathit{it}}(\mathbf{x}_i^t)) \in T_u^{\mathit{it}+1} : -1 \leq f^{\mathit{it}}(\mathbf{x}_i^t) < 0 \right\}$ (par ordre décroissant selon la marge)

$S^{\mathit{it}+1} \leftarrow S^{\mathit{it}} \bigcup \tau$ premiers éléments de $H_{up}^{\mathit{it}} \bigcup \tau$ premiers éléments de H_{low}^{it}

$L \leftarrow \left\{ (\mathbf{x}^t, h^{\mathit{it}}(\mathbf{x}_i^t)) \in S^{\mathit{it}} \bigcap T_u^1 : h^{\mathit{it}}(\mathbf{x}_i^t) \neq h_{\mathit{it}-1}(\mathbf{x}_i^t) \right\}$

$T_u^{\mathit{it}+1} \leftarrow T_u^{\mathit{it}+1} \bigcup L$

$S^{\mathit{it}+1} \leftarrow S^{\mathit{it}+1} \bigcup S^{\mathit{it}} \backslash L$

$Q_{up}^{\mathit{it}} \leftarrow \{ (\mathbf{x}^s, y^s) \in S^{\mathit{it}} \bigcap S^1 : f_{\mathit{it}}(\mathbf{x}^s) \geq 0 \}$ (par ordre décroissant selon la marge)

$Q_{low}^{\mathit{it}} \leftarrow \{ (\mathbf{x}^s, y^s) \in S^{\mathit{it}} \bigcap S^1 : f_{\mathit{it}}(\mathbf{x}^s) < 0 \}$ (par ordre décroissant selon la marge)

$S^{\mathit{it}+1} \leftarrow S^{\mathit{it}+1} \backslash k$ premiers éléments de Q_{up}^{it}, tel que k corresponde au nombre d'exemples de H_{up}^{it} ajoutés à $S^{\mathit{it}+1}$ (s'il vaut 0 alors $k = \tau$)

$S^{\mathit{it}+1} \leftarrow S^{\mathit{it}+1} \backslash k$ premiers éléments de Q_{low}^{it}, tel que k corresponde au nombre d'exemples de H_{low}^{it} ajoutés à $S^{\mathit{it}+1}$ (s'il vaut 0 alors $k = \tau$)

jusqu'à $|Q_{up}^{\mathit{it}} \bigcup Q_{low}^{\mathit{it}}| = 0$ OU $|L| < \beta m^s$ OU $|H_{up}^{\mathit{it}} \bigcup H_{low}^{\mathit{it}}| \leq \beta m^s$

retourner h^{it}

Nous présentons l'algorithme d'adaptation de domaine appelé DASVM [Bruzzone et Marconcini, 2010]. Cet algorithme adapte itérativement le classifieur-SVM appris à l'aide de données cibles auto-étiquetées. Concrètement, à chaque itération it, DASVM apprend un classifieur-SVM à partir des exemples étiquetés S^{it}. Il étiquette ensuite les exemples de l'ensemble cible T_u^{it} et en rajoute un certain nombre dans l'échantillon étiqueté S^{it}. Ces données cibles auto-étiquetés correspondent à celles dont la confiance du classifieur est proche et inférieure à la marge. En outre, il retire de l'échantillon étiqueté S^{it} les exemples dont la confiance est la plus importante. Le processus est réitéré avec le nouvel ensemble étiqueté $S^{\mathit{it}+1}$ obtenu et se termine lorsque tous les exemples cibles T_u ont été ajoutés à l'échantillon étiqueté. DASVM est illustré sur la figure 2.5 et décrit dans l'algorithme 2.

Il est important de remarquer que cet algorithme ne se base pas sur la théorie présentée précédemment. Il existe peu de garanties théoriques pour cet approche, hormis les travaux de [Habrard *et al.*, 2011] qui donnent des garanties minimalistes. Cependant, en pratique, DASVM montre de bons résultats, c'est pourquoi nous ferons appel à lui en tant que méthode de référence.

2.3.2 CODA : un algorithme d'adaptation par co-apprentissage

CODA [Chen *et al.*, 2011a] est un algorithme dit de co-apprentissage pour l'adaptation de domaine et est une variante de l'algorithme PMC (*Pseudo-Multiview Co-training* [Chen *et al.*, 2011b]). Il essaie de réduire la divergence entre les domaines en ajoutant itérativement, à l'ensemble d'apprentissage, des attributs cibles et des exemples pour lesquels l'algorithme est le plus confiant : à chaque itération, CODA résoud un problème d'optimisation simple en apprenant simultanément une hypothèse cible, une division de l'espace en deux descriptions/vues différentes et un sous-ensemble d'attributs sources et cibles à inclure dans l'hypothèse. Concrètement, le co-apprentissage requiert les points suivants.

- Deux classifieurs qui, ici, sont des classifieurs linéaires noté $h_{\mathbf{u}}$ et $h_{\mathbf{v}}$ identifiés par les vecteurs de poids $\mathbf{u} = (u_1, \ldots, u_d)^\top$ et $\mathbf{v} = (v_1, \ldots, v_d)^\top$. La performance de chaque classifieur est mesurée via la fonction de perte logistique $\ell_{\log}(\cdot, \cdot)$ définie pour tout classifieur linéaire $h_{\mathbf{w}}$ identifié par le vecteur $\mathbf{w} \in \mathbb{R}^d$ par :

$$\forall \mathbf{x} \in \mathbb{R}^d,\ \ell(h_{\mathbf{w}}(\mathbf{x}), y) = \log\left(1 + \exp\left(-y\langle \mathbf{w}, \mathbf{x}\rangle\right)\right).$$

  Afin de s'assurer que les deux classifieurs $h_{\mathbf{u}}$ et $h_{\mathbf{v}}$ ont tous les deux un risque faible sur l'échantillon d'apprentissage étiqueté S, ils sont appris conjointement en minimisant le *soft-maximum* des deux risques empiriques :

$$\log\left[\exp\left(\mathbf{R}_S^\ell(h_{\mathbf{u}})\right) + \exp\left(\mathbf{R}_S^\ell(h_{\mathbf{v}})\right)\right].$$

- Les deux classifieurs doivent être appris sur deux vues différentes. Elles sont créées en divisant la représentation originelle en deux sous-espace exclusifs. Plus précisément, pour chaque attribut identifié par l'indice i, au moins un des deux classifieur doit admettre un poids nul pour la $i^{\text{ième}}$ dimension. Pour ce faire, il suffit de contraindre la résolution du problème d'optimisation de sorte que :

$$\sum_{i=1}^{d} u_i^2 v_i^2 = 0.$$

- Dans la formulation d'origine du co-apprentissage, les deux vues doivent être indépendantes sachant la classe. Pour relâcher cette restriction, CODA vérifie la condition de *ϵ-expandability* [Balcan *et al.*, 2004]. Intuitivement, deux classifieurs capables d'apprendre l'un de l'autre doivent être confiants sur deux sous-ensembles différents de données non étiquetées. Plus formellement, on note la confiance d'un classifieur linéaire $\mathbf{w}$ pour un exemple $\mathbf{x}$ par :

$$c_{\mathbf{w}}(\mathbf{x}) = \begin{cases} 1 & \text{si la confiance dépasse un certain seuil,} \\ 0 & \text{sinon.} \end{cases}$$

  Étant donnés $\epsilon > 0$, un échantillon non étiqueté U, la condition de *ϵ-expandability* est alors définie par :

$$\sum_{\mathbf{x} \in U} \left[c_{\mathbf{u}}(\mathbf{x})\bar{c}_{\mathbf{v}}(\mathbf{x}) + \bar{c}_{\mathbf{u}}(\mathbf{x})c_{\mathbf{v}}(\mathbf{x})\right] \geq \epsilon \min\left\{\sum_{\mathbf{x} \in U} \bar{c}_{\mathbf{u}}\bar{c}_{\mathbf{v}}(\mathbf{x})\right\},$$

où $\bar{c}_{\mathbf{w}} = 1 - c_{\mathbf{w}}(\mathbf{x})$.

Le classifieur linéaire final appris est celui associé au vecteur de poids : $\mathbf{w} = \mathbf{u} + \mathbf{v}$. De plus, pour résoudre le problème les auteurs ont ajouté une régularisation, notée $\text{reg}(S, U, \mathbf{w})$ et basée sur la norme 1, pour induire une parcimonie sur les attributs dont les corrélations sont opposées ou faibles sur les domaines. Le problème d'optimisation à résoudre est alors le suivant :

$$\begin{cases} \min\limits_{\mathbf{w},\mathbf{u},\mathbf{v}} & \log\left(\exp\left(\mathbf{R}_S^{\ell}(\mathbf{u})\right) + \exp\left(\mathbf{R}_S^{\ell}(\mathbf{v})\right)\right) + \text{reg}(S, U, \mathbf{w}), \\ \text{s.c.} & \sum\limits_{i=1}^{d} u_i^2 v_i^2 = 0, \\ & \sum\limits_{\mathbf{x}\in U} \left[c_{\mathbf{u}}(\mathbf{x})\bar{c}_{\mathbf{v}}(\mathbf{x}) + \bar{c}_{\mathbf{u}}(\mathbf{x})c_{\mathbf{v}}(\mathbf{x})\right] \geq \epsilon \min\left\{\sum\limits_{\mathbf{x}\in U} \bar{c}_{\mathbf{u}}\bar{c}_{\mathbf{v}}(\mathbf{x})\right\}, \\ & \mathbf{w} = \mathbf{u} + \mathbf{v}. \end{cases}$$

Ce problème d'optimisation n'est pas convexe, cependant, il n'est pas sensible à son initialisation [Chen *et al.*, 2011a]. Ainsi, il suffit de fixer aléatoirement $\mathbf{u}$ et $\mathbf{v}$ puis de l'optimiser avec une descente de gradient conjugué standard.

2.3.3 Validation des hyperparamètres

Quelle que soit l'approche que l'on choisit de suivre, une problématique encore ouverte en adaptation de domaine est la sélection ou la validation des différents hyperparamètres des algorithmes. En effet, puisque nous supposons que le domaine source et le domaine cible sont différents et, qu'en plus, nous avons peu ou pas d'étiquette(s) cible(s), les processus de validation usuels, tels que la validation croisée, ne peuvent être appliqués. Notons que les récents travaux de [Geras et Sutton, 2013] proposent une procédure de validation lorsque plusieurs domaines sources sont disponibles. Dans cette thèse, nous allons faire appel à une validation dite circulaire [Bruzzone et Marconcini, 2010] ou inverse [Zhong *et al.*, 2010]. Étant donnés k sous-ensembles [12] de l'échantillon étiqueté source S et un algorithme d'adaptation de domaine, nous en utilisons $k-1$ en tant qu'exemples d'apprentissage pour apprendre un classifieur h avec cet algorithme. On étiquette ensuite l'échantillon cible non étiqueté T_u à l'aide de h et on obtient un échantillon auto-étiqueté $\widehat{T_u} = \{(\mathbf{x}_i^t, h(\mathbf{x}_i^t))\}_{i=1}^{m_u^t}$. Puis, on ré-applique l'algorithme d'adaptation avec S_u un échantillon source non étiqueté, $\widehat{T_u}$ et l'échantillon cible étiqueté T, lorsqu'il est disponible, pour apprendre le classifieur inverse h^r. En fait, on inverse le rôle du domaine source et du domaine cible. Le classifieur inverse h^r est alors évalué sur le dernier $k^{\text{ème}}$ sous-ensemble issu de l'échantillon source S avec l'intuition que si l'adaptation est possible, alors on doit pouvoir "facilement" passer d'un domaine à l'autre grâce à l'adaptation. Nous illustrons ce principe sur la figure 2.6.

12. On parle de *k-folds* en anglais.

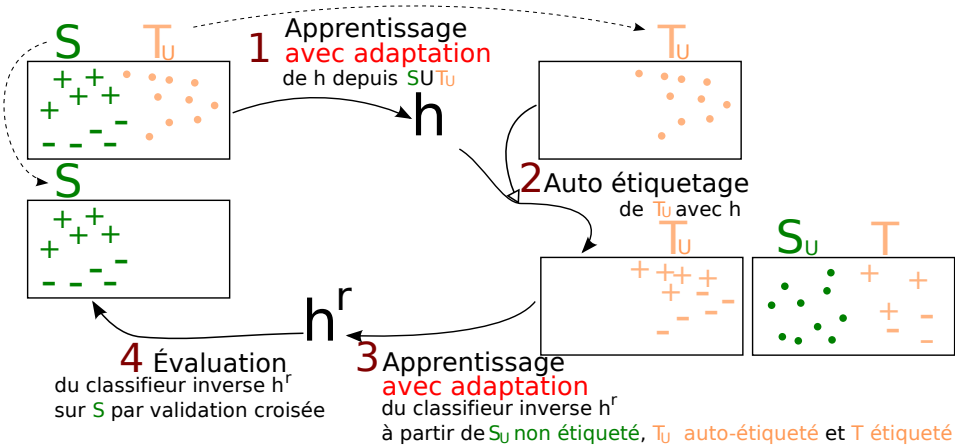


FIGURE 2.6 – *Le processus de validation inverse. Étape* **1** *: Apprendre avec adaptation le classifieur h. Étape* **2** *: Auto-étiqueter l'échantillon cible non étiqueté avec h. Étape* **3** *: Apprendre avec le même algorithme d'adaptation le classifieur inverse h^r à partir de l'échantillon source non étiqueté, de l'échantillon cible auto-étiqueté et de l'échantillon cible étiqueté disponible. Étape* **4** *: Évaluer h^r sur l'échantillon source étiqueté.*

2.4 SYNTHÈSE

L'adaptation de domaine est une tâche complexe. En effet, si aucune information sur les étiquettes cibles n'est disponible, il est nécessaire de supposer qu'il existe une relation entre les domaines. Pour s'attaquer à ce problème, une solution est alors de considérer l'hypothèse qu'il existe une relation forte entre les domaines. Un cas particulier est le *covariate shift* pour lequel on suppose que la probabilité d'étiqueter un exemple par une classe sachant l'exemple est la même. Cependant, ce cas reste idylique et de nombreux algorithmes s'inspirent de l'analyse classique qui se base sur une notion de divergence entre les domaines, pour estimer dans quelle mesure nous sommes capables d'adapter. Le principe général d'un algorithme d'adaptation est alors d'apprendre un classifieur performant tout en rapprochant les marginales selon cette divergence (par exemple en projetant les données dans un espace commun). Nous proposerons dans le chapitre 6 un algorithme basé sur cette théorie en cherchant à rapprocher les domaines dans l'espace de projection explicite défini par une fonction de similarité (ϵ,γ,τ)-bonne (dont le principe en apprentissage supervisé a été énoncé en section 1.4.3). Néanmoins, le cadre classique présenté en section 2.2 montre des inconvénients puisque la divergence considérée peut être vue comme une analyse dans le pire cas : il est difficile d'optimiser en même temps la divergence et l'erreur source. Nous verrons en chapitre 7 comment faire appel à la théorie PAC-Bayésienne pour contourner les inconvénients portés par l'analyse classique.

La problématique générale de cette thèse repose sur l'apprentissage de vote de majorité. Nous présentons donc, dans le chapitre suivant, la théorie PAC-Bayésienne en classification supervisée qui, au contraire des approches précédemment présentées, se focalise sur l'apprentissage de votes de majorité pondérés sur un ensemble de votants tels que des classifieurs ou des fonctions plus générales.

Théorie PAC-Bayésienne et vote de majorité

3

Dans ce chapitre, nous nous replaçons dans le cadre de l'apprentissage supervisé sans adaptation et nous nous focalisons sur la théorie PAC-Bayésienne [1], introduite par [McAllester, 1999]. Cette théorie, qui fournit un champ de recherche important en théorie de l'apprentissage, a pour but premier d'offrir des bornes en généralisation sur l'erreur de votes de majorité pondérés sur une famille d'hypothèses $\mathcal{H}$. Plus précisément, les ingrédients d'une borne PAC-Bayésienne sont :

- un échantillon d'apprentissage S dont les éléments sont *i.i.d.* selon un domaine P sur $X \times Y$;
- une distribution *a priori* π sur $\mathcal{H}$ qui modélise une certaine connaissance *a priori*, c'est-à-dire avant l'observation de S, sur la performance des hypothèses de $\mathcal{H}$: celles supposées les plus performantes, pour une certaine tâche, verront leur probabilité selon π plus élevée ;

1. Le lecteur peut se référer au tutoriel ICML 2012 : "*PAC-Bayesian Analysis in Supervised, Unsupervised, and Reinforcement*" (`people.kyb.tuebingen.mpg.de/seldin/ICML_Tutorial_PAC_Bayes.htm`) de Yevgeny Seldin, François Laviolette et John Shawe-Taylor.

- une distribution *a posteriori* ρ sur $\mathcal{H}$ qui est apprise ou ajustée à la lumière de l'information apportée par l'échantillon d'apprentissage S. Cette distribution permet de définir le vote de majorité pondéré : le poids d'un votant correspond à sa probabilité d'apparaître selon ρ.

En fait, la théorie PAC-Bayésienne tire son inspiration de la philosophie de l'inférence bayésienne mixée à des techniques d'apprentissage statistique. En effet, l'inférence bayésienne [2] suppose une distribution *a priori* sur $\mathcal{H}$, puis fait appel à la règle de Bayes pour inférer la distribution *a posteriori* en se basant sur la vraissemblance des données pour chaque hypothèse. Les capacités en généralisation en inférence bayésienne présument de la correcte définition de la distribution *a priori*, tandis qu'une borne en généralisation PAC-Bayésienne est vraie pour n'importe quel choix de distribution *a priori*. Concrètement, l'analyse PAC-Bayésienne étudie l'espérance, selon la distribution ρ, des erreurs des votants. En général, elle étudie l'erreur du classifieur stochastique de Gibbs : il prédit l'étiquette d'un exemple $\mathbf{x}$ en tirant aléatoirement selon ρ une hypothèse h dans l'ensemble $\mathcal{H}$, puis en retournant $h(\mathbf{x})$. La complexité de la classe d'hypothèses est ici implicitement capturée par la divergence de Kullback-Leibler (notée KL-divergence) entre les distributions *a posteriori* et *a priori* : elle permet de cconsidérer une information sur la complexité de chaque hypothèse contrairement aux approches classiques. Finalement, si la distribution apprise ρ affecte une probabilité élevée aux hypothèses de $\mathcal{H}$ suffisamment performantes et si ρ et π sont relativement proches au sens de la KL-divergence, alors la borne sur l'erreur du classifieur de Gibbs peut se montrer informative et précise pour le vote de majorité sur $\mathcal{H}$ pondéré par ρ.
De plus, en la spécialisant à des espaces $\mathcal{H}$ appropriés, à des familles de distributions *a priori* et *a posteriori* adéquates ou à des fonctions de perte spécifiques, l'approche PAC-Bayésienne permet de caractériser les capacités en généralisation de méthodes existantes. Nous pouvons citer en exemple, la borne PAC-Bayésienne prouvée pour les SVM par [Langford et Shawe-Taylor, 2002] lorsque les distributions *a priori* et *a posteriori* suivent toutes deux une distribution normale (de moyenne et variance différentes). En outre, l'étude empirique de [Langford, 2005] a révélé que cette borne est un estimateur précis du risque des SVM que l'on peut encore améliorer en prenant en compte une distribution *a priori* informative [3] [Ambroladze *et al.*, 2006, Parrado-Hernández *et al.*, 2012]. Une telle analyse permet d'obtenir de meilleures garanties que les analyses classiques. Signalons que les travaux de [Blanchard et Fleuret, 2007], avec le *Occam's Hammer*, proposent une alternative intéressante à la théorie PAC-Bayésienne, mais n'entrent pas dans le champ d'étude de ce manuscrit.

Dans ce chapitre, nous introduisons tout d'abord, en section 3.1, les notions de vote de majorité et de classifieur de Gibbs, ainsi que les liens qui les unissent. Puis nous énonçons en section 3.2 les bornes PAC-Bayésiennes classiques. Nous présentons, ensuite, deux algorithmes basés sur la théorie PAC-Bayésienne. Le premier, PBGD

2. Pour plus d'information sur l'inférence bayésienne, qui tire son nom du théorème de Bayes, le lecteur peut se référer à [Gelman *et al.*, 2004].

3. Une distribution non-informative correspond, par exemple, à la distribution uniforme.

[Germain *et al.*, 2009a] en section 3.3, optimise l'erreur du classifieur de Gibbs sur un ensemble de classifieurs linéaires en minimisant une des bornes PAC-Bayésiennes. Le second, MinCq [Laviolette *et al.*, 2011a] en section 3.4, optimise directement l'erreur du vote majorité en minimisant une borne qui relie étroitement le classifieur de Gibbs et le vote de majorité.

3.1 Vote de majorité et classifieur stochastique de Gibbs

Nous nous plaçons dans le cadre classique de la théorie PAC-Bayésienne en classification supervisée binaire avec $Y = \{-1, +1\}$. Traditionnellement, on considère les votes de majorité pondérés construits à partir d'un ensemble d'hypothèses $\mathcal{H}$ que l'on appelle les votants. Étant donnés une distribution *a priori* π sur $\mathcal{H}$ (que nous appelons la distribution prior ou le prior) et un échantillon d'apprentissage S dont les éléments sont tirés *i.i.d.* selon P, l'apprenant doit trouver la distribution *a posteriori* ρ sur $\mathcal{H}$ (que nous appelons la distribution posterior ou le posterior) amenant au vote de majorité ρ-pondéré $B_\rho(\cdot)$ offrant les meilleures garanties en généralisation.

Définition 3.1 (Vote de majorité ρ-pondéré) *Soit $\mathcal{H} = \{h_1, \ldots, h_n\}$ un ensemble de n votants de X vers Y. Soit ρ une distribution sur $\mathcal{H}$. Le vote de majorité ρ-pondéré (parfois appelé classifieur de Bayes dans la litterature) $B_\rho(\cdot)$ associé à ρ est défini par :*

$$\begin{aligned} \forall \mathbf{x} \in X,\ B_\rho(\mathbf{x}) &= \operatorname{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right] \\ &= \operatorname{sign}\left[\sum_{h\in\mathcal{H}} \rho(h) h(\mathbf{x})\right]. \end{aligned} \tag{3.1}$$

Son erreur réelle $\mathbf{R}_P(B_\rho)$ sur un domaine P sur $X \times Y$ est :

$$\begin{aligned} \mathbf{R}_P(B_\rho) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \ell_{0\text{–}1}\left(B_\rho(\mathbf{x}), y\right) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathbf{I}\left(B_\rho(\mathbf{x}) \neq y\right). \end{aligned}$$

L'erreur empirique estimée sur un échantillon S dont les éléments sont tirés i.i.d. selon P est :

$$\mathbf{R}_S(B_\rho) = \frac{1}{m} \sum_{(\mathbf{x},y)\in S} \ell_{0\text{–}1}\left(B_\rho(\mathbf{x}), y\right).$$

L'objectif est donc de trouver la distribution posterior ρ telle que l'erreur réelle $\mathbf{R}_P(B_\rho)$ soit la plus faible possible. Cependant, la minimisation de cette erreur est connue pour être NP-dure. Habituellement, $\mathbf{R}_P(B_\rho)$ est alors "remplacée" par l'espérance selon ρ des erreurs des votants de $\mathcal{H}$: $\mathbf{E}_{h\sim\rho}\, \mathbf{R}_P(h)$. Ce moyennage des erreurs correspond, en fait, à l'erreur du classifieur stochastique de Gibbs $G_\rho(\cdot)$ associé à la distribution posterior ρ. Pour prédire l'étiquette d'un exemple $\mathbf{x}$, $G_\rho(\mathbf{x})$ tire aléatoirement un votant h dans $\mathcal{H}$ selon ρ, puis renvoie $h(\mathbf{x})$. L'erreur du classifieur de Gibbs est donc définie par :

$$\mathbf{R}_P(G_\rho) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_P(h).$$

D'après cette définition, il existe une relation triviale entre $B_\rho(\cdot)$ et $G_\rho(\cdot)$. En effet, si $B_\rho(\cdot)$ se trompe pour un exemple $\mathbf{x}$, alors au moins la moitié des classifieurs selon ρ commettent une erreur sur $\mathbf{x}$. De ce fait, les erreurs $\mathbf{R}_P(B_\rho)$ et $\mathbf{R}_P(G_\rho)$ sont liées par la relation suivante.

Proposition 3.1 *Pour tout domaine P sur $X \times Y$ et pour toute distribution ρ sur $\mathcal{H}$, on a :*

$$\mathbf{R}_P(B_\rho) \leq 2\mathbf{R}_P(G_\rho).$$

D'après cette relation, une borne en généralisation sur l'erreur du classifieur de Gibbs implique une majoration de l'erreur du vote de majorité à un facteur 2 près. Nous verrons dans le chapitre 5 que ce facteur correspond, en fait, au nombre de classes considérées. Dans certaines situations, ce type de borne s'avère précise : par exemple, lorsque le classifieur de Gibbs admet une erreur faible, autrement dit que les erreurs individuelles sont, en moyenne, faibles. Cependant, ce facteur peut amener à une borne supérieure à 1 : lorsque les erreurs sont en moyenne supérieures à 0.5. Dans le cas de classifieurs ayant une grande marge de séparation, ce facteur peut néanmoins être réduit à $(1+\epsilon)$ (pour un petit ϵ positif) [Langford et Shawe-Taylor, 2002], mais cette relation souffre des même problèmes que la précédente : la situation $\mathbf{R}_P(B_\rho) \ll \mathbf{R}_P(G_\rho)$ reste fréquente.
Toutefois, une relation — bien plus précise — existe entre $\mathbf{R}_P(G_\rho)$ et $\mathbf{R}_P(B_\rho)$: la C-borne [Lacasse *et al.*, 2007]. Elle lie l'erreur du vote de majorité à la moyenne et la variance de l'erreur du classifieur de Gibbs qui correspondent respectivement à l'espérance des erreurs et le désaccord/la diversité des votants selon ρ. En ce sens, la borne est plus informative mais plus complexe à manipuler que la relation triviale de la proposition 3.1.

Théorème 3.1 (La C-borne de [Lacasse *et al.*, 2007]) *Pour toute distribution ρ sur un ensemble de votants $\mathcal{H}$, et pour tout domaine P sur $X \times Y$, si $\mathbf{R}_P(G_\rho) \leq \frac{1}{2}$, alors on a :*

$$\mathbf{R}_P(B_\rho) \leq C_P^\rho,$$

avec :

$$C_P^\rho = 1 - \frac{\left(1 - 2\mathbf{R}_P(G_\rho)\right)^2}{1 - 2\mathbf{R}_D(G_\rho, G_\rho)},$$

où $\mathbf{R}_D(G_\rho, G_\rho)$ correspond aux désaccords des votants selon ρ que l'on définit par :

$$\begin{aligned}\mathbf{R}_D(G_\rho, G_\rho) &= \mathop{\mathbf{E}}_{h,h' \sim \rho^2} \mathbf{R}_D(h, h') \\ &= \mathop{\mathbf{E}}_{h,h' \sim \rho^2} \mathop{\mathbf{E}}_{\mathbf{x} \sim D} \mathbf{I}\left(h(\mathbf{x}) \neq h'(\mathbf{x})\right).\end{aligned}$$

Démonstration. Voir en annexe B.3. □

Nous reprendrons ce résultat en section 3.4, puis nous présenterons l'algorithme MinCq [Laviolette *et al.*, 2011a] qui en découle et qui permet d'apprendre un vote de majorité ρ-pondéré performant sur un ensemble de votants à valeurs réelles.

Sachant que le vote de majorité et le classifieur de Gibbs sont étroitement liés, borner l'espérance des erreurs $\mathbf{E}_{h\sim\rho}\,\mathbf{R}_P(h) = \mathbf{R}_P(G_\rho)$ apparaît pertinent. Ainsi, les bornes PAC-Bayésiennes s'expriment en fonction de deux quantités principales : l'erreur empirique du classifieur de Gibbs $\mathbf{R}_S(G_\rho)$ estimée sur l'échantillon d'apprentissage $S \sim (P)^m$ et la KL-divergence entre la distribution posterior apprise ρ et la distribution prior π qui est définie par :

$$\mathrm{KL}(\rho\|\pi) = \mathop{\mathbf{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}. \tag{3.2}$$

Nous pouvons maintenant introduire la formulation générique du théorème PAC-Bayes dans sa version la plus simple, c'est-à-dire celle s'appliquant à l'erreur standard mesurée par la fonction de perte $0-1$. Notons qu'en suivant le même principe de preuve, ce résultat peut-être étendu à des mesures de risque plus complexes.

3.2 Le théorème PAC-Bayes

3.2.1 Un théorème qui englobe les autres

Dans la littérature, le théorème PAC-Bayes se décline sous différentes versions en fonction de la mesure considérée pour comparer l'erreur réelle et son estimation empirique. Les principales sont dues à [McAllester, 1999, McAllester, 2003, Seeger, 2002, Langford, 2005, Catoni, 2007]. Cependant, [Germain *et al.*, 2009a] ont proposé la formulation générale suivante permettant de retrouver simplement ces versions déjà connues, mais aussi de dériver de nouvelles bornes.

Théorème 3.2 (Théorème PAC-Bayes [Germain *et al.*, 2009a]) *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de votants, soit S un échantillon dont les éléments sont tirés i.i.d. selon P. Alors pour toute distribution prior π sur $\mathcal{H}$, pour toute fonction convexe $\mathcal{D} : [0,1] \times [0,1] \mapsto \mathbb{R}$ et pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, pour toute distribution posterior ρ sur $\mathcal{H}$, on a :*

$$\mathcal{D}\left(\mathbf{R}_S(G_\rho), \mathbf{R}_P(G_\rho)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}\right)\right].$$

Démonstration. Voir en annexe B.1. □

Ce théorème est valide pour toutes les distributions posterior ρ et, en particulier, pour celle associée au meilleur vote de majorité ρ-pondéré. Il exprime les garanties en généralisation en majorant l'écart entre la valeur réelle $\mathbf{R}_P(G_\rho)$ et son estimateur $\mathbf{R}_S(G_\rho)$. Cet écart se mesure via une fonction de comparaison $\mathcal{D}(\cdot,\cdot)$. En outre, cette borne dépend de la KL-divergence entre le posterior et le prior, du nombre d'exemples d'apprentissage m ainsi que de l'espérance pour tous les votants selon π et pour tous les échantillons aléatoires de taille m : $\mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}$. Ainsi, pour dériver une borne en généralisation, il suffit de définir une fonction convexe $\mathcal{D} : [0,1] \times [0,1] \mapsto \mathbb{R}$, puis de calculer ou de majorer (avec par exemple une inégalité de concentration) :

$\mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} e^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}$. C'est en suivant ce principe simple que l'on retrouve les bonnes classiques énoncées dans la section 3.2.3 en considérant des fonctions $\mathcal{D}(\cdot,\cdot)$ différentes.

Enfin, soulignons que le théorème 3.2 a été étendu à des fonctions de pertes plus générales [Germain *et al.*, 2009b] (autre qu'une fonction de perte linéaire en ρ). Nous en verrons une illustration en section 3.4, où la perte considérée est liée à la notion de marge du vote de majorité.

3.2.2 Quelques mots sur la philosophie de la théorie PAC-Bayésienne

À l'instar des bornes en généralisation présentées en section 1.3, l'approche PAC-Bayésienne exprime des garanties vérifiables avec une grande probabilité et qui dépendent de la fonction de perte considérée. À la différence des analyses VC et de Rademacher (sections 1.3.1 et 1.3.2), la complexité est ici définie pour chaque votant en fonction de leur probabilité *a priori* $\pi(h)$ plutôt qu'une complexité globale sur l'ensemble des votants. En ce sens, la borne considère les votants tirés aléatoirement selon la distribution posterior ρ (on parle parfois de borne "en moyenne") et dépend donc de la divergence entre les "complexités" individuelles $\rho(h)$ et $\pi(h)$ capturées par $\mathrm{KL}(\rho\|\pi)$. Au contraire des approches classiques, mais à l'image de l'inférence bayésienne, cette divergence permet d'incorporer explicitement une connaissance *a priori* sur chacun des votants : plus le posterior est proche du prior, plus la borne sera précise. Définir une "bonne" distribution prior est donc un problème pertinent et important en théorie PAC-Bayésienne. Dans la littérature, différentes approches existent pour apprendre le prior à partir d'un sous-échantillon des données afin d'obtenir des bornes plus précises [4] [Ambroladze *et al.*, 2006, Germain *et al.*, 2009a, Parrado-Hernández *et al.*, 2012]. Par ailleurs, bien que le domaine P soit inconnu, il est possible, dans certaine situation, d'estimer la valeur de $\mathrm{KL}(\rho\|\pi)$ [Catoni, 2003, Langford, 2005, Lever *et al.*, 2010, Lever *et al.*, 2013]. Nous en verrons une illustration en section 3.3. De plus, lorsque le prior reste complexe à calculer, une astuce élégante [Laviolette *et al.*, 2011a], présentée en section 3.4, permet de s'affranchir de la mesure de complexité, simplifiant alors la dérivation d'algorithmes. Enfin, nous verrons dans le chapitre 4 que cette astuce peut être combinée à un *a priori* informatif à des fins algorithmiques. Ces approches permettent d'obtenir des bornes parfois très précises en limitant ou supprimant l'importance du terme de complexité.

Signalons que les résultats classiques sont valides, d'une part, lorsque les données sont indépendemment et identiquement distribuées selon P et, d'autre part, lorsque les votants sont indépendants des données d'apprentissage. Si les données ne sont pas *i.i.d.* selon P, nous pouvons nous référer aux travaux de [Ralaivola *et al.*, 2010, Lever *et al.*, 2010, Lever *et al.*, 2013]. Si les votants dépendent des données, une solution est de faire appel aux résultats proposés par [Graepel *et al.*, 2005, Laviolette et Marchand, 2007, Germain *et al.*, 2011] qui

4. Puisque minimiser $\mathrm{KL}(\rho\|\pi)$ revient à rapprocher le posterior du prior au sens de la KL-divergence.

généralisent la théorie PAC-Bayésienne aux schémas de compression (voir le chapitre 4 pour un exemple de borne PAC-Bayésienne dans une telle situation). Finalement, la théorie PAC-Bayésienne démontre aussi son intérêt dans de nombreux problèmes d'apprentissage automatique, tels que la classification à sortie structurée [Giguère *et al.*, 2013], le clustering [Seldin et Tishby, 2010, Higgs et Shawe-Taylor, 2010] ou l'apprentissage par renforcement [Seldin *et al.*, 2012].

3.2.3 Les bornes PAC-Bayésiennes classiques

Nous détaillons maintenant comment retrouver les trois versions classiques du théorème PAC-Bayes à partir du théorème général 3.2. Concrètement, il suffit d'appliquer ce théorème à des fonctions de comparaison $\mathcal{D}(\cdot,\cdot)$ différentes.

Tout d'abord, si l'on considère la KL-divergence entre des distributions de Bernoulli avec une probabilité de succès a et d'échec b, on obtient une borne un peu plus précise et similaire à celle proposée par [Seeger, 2002, Langford, 2005]. Ainsi, en appliquant le théorème 3.2 avec $\mathcal{D}(a,b) = \mathrm{kl}(a\|b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$, on a :

Corollaire 3.1 ([Seeger, 2002, Langford, 2005]) *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de votants, soit S un échantillon de m éléments tirés i.i.d. selon P. Alors pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, pour toute distribution posterior ρ sur $\mathcal{H}$, on a :*

$$\mathrm{kl}\left(\mathbf{R}_S(G_\rho), \mathbf{R}_P(G_\rho)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}\right],$$

où $\xi(m) = \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k} \leq m+1.$

La version de [Seeger, 2002, Langford, 2005] est obtenue en remplaçant $\xi(m)$ par sa majoration $m+1$.

Démonstration du corollaire 3.1.[5] On applique le théorème 3.2 avec $\mathcal{D}(a,b) = \mathrm{kl}(a,b)$, puis on majore $\mathop{\mathbf{E}}_{S\sim(P)^m}\mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim(P)^m}\mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))} &= \mathop{\mathbf{E}}_{h\sim\pi}\mathop{\mathbf{E}}_{S\sim(P)^m}\left(\frac{\mathbf{R}_S(h)}{\mathbf{R}_P(h)}\right)^{m\mathbf{R}_S(h)}\left(\frac{1-\mathbf{R}_S(h)}{1-\mathbf{R}_P(h)}\right)^{m(1-\mathbf{R}_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim\pi}\sum_{k=0}^{m}\mathop{\mathbf{Pr}}_{S\sim(P)^m}\left(\mathbf{R}_S(h)=\frac{k}{m}\right)\left(\frac{k/m}{\mathbf{R}_P(h)}\right)^k\left(\frac{1-k/m}{1-\mathbf{R}_P(h)}\right)^{m-k} \\
&= \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k} = \xi(m) \\
&\leq m+1.
\end{aligned}$$

La dernière ligne est obtenue car :

$$\xi(m) \leq \mathop{\mathbf{E}}_{h\sim\pi}\sum_{i=0}^{m} 1 = \mathop{\mathbf{E}}_{h\sim\pi}(m+1) = m+1.$$

5. Une preuve similaire a été proposée par [Banerjee, 2006].

□

Cette version du théorème PAC-Bayes offre une borne précise, en particulier pour des erreurs empiriques faibles, mais néanmoins difficile à interpréter à cause du terme : $\mathrm{kl}\left(\mathbf{R}_S(G_\rho), \mathbf{R}_P(G_\rho)\right)$. En effet, contrairement aux bornes en généralisation présentées dans la section 1.3, $\mathbf{R}_P(G_\rho)$ n'est pas explicitement bornée par l'erreur empirique $\mathbf{R}_S(G_\rho)$. Son optimisation à des fins algorithmiques n'apparaît donc pas aisée en général. En ce sens, la version "historique" de [McAllester, 1999], obtenue en considérant la relation linéaire : $\mathcal{D}(a,b) = 2(a-b)^2$, s'interprète plus facilement car elle borne explicitement l'écart entre les erreurs réelle et empirique, mais s'avère moins précise.

Corollaire 3.2 *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de votants, soit S un échantillon de m éléments tirés i.i.d. selon P.. Alors pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, pour toute distribution posterior ρ sur $\mathcal{H}$, on a :*

$$\left(\mathbf{R}_S(G_\rho) - \mathbf{R}_P(G_\rho)\right)^2 \leq \frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}\right],$$

où $\xi(m) = \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k} \leq m+1.$

Démonstration. La borne est dérivée du corollaire 3.1 grâce à l'inégalité de Pinkser : $2(a-b)^2 \leq \mathrm{kl}(a\|b)$. □

Notons qu'il existe également une preuve "alternative", en trois étapes, de cette borne qui se base sur l'inégalité de Hoeffding [6] [McAllester, 2003] et que nous utiliserons dans le chapitre 5.
Que ce soit le corollaire 3.1 ou le corollaire 3.2, le compromis entre la KL-divergence et le risque empirique ne peut être contrôlé explicitement. L'approche proposée par [Catoni, 2007] est hyperparamétrée et permet ce contrôle. Soit la fonction convexe $\mathcal{F}(b) = \ln\frac{1}{1-b[1-\exp(-C)]}$, cette borne se dérive en posant $\mathcal{D}(a,b) = \mathcal{F}(b) - Ca$, avec $C > 0$ un paramètre de régularisation.

Corollaire 3.3 ([Catoni, 2007]) *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de votants, soit S un échantillon de m éléments tirés i.i.d. selon P. Alors pour toute distribution prior π sur $\mathcal{H}$, pour tout réel $C > 0$ et pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, pour toute distribution posterior ρ sur $\mathcal{H}$, on a :*

$$\mathbf{R}_P(G_\rho) \leq \frac{1}{1-\mathrm{e}^{-C}}\left\{1 - \exp\left(-\left[C\mathbf{R}_S(G_\rho) + \frac{1}{m}\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}\right)\right]\right)\right\}. \tag{3.3}$$

Démonstration. Voir en annexe B.2. □

Si $C = \frac{1}{\sqrt{m}}$, cette borne devient consistante : lorsque m tend vers $+\infty$, la borne tend vers $1 \times \left[\mathbf{R}_S(G_\rho) + 0\right]$.

6. L'inégalité de Hoeffding est énoncée dans le théorème A.1 en annexe A.

Similairement à la borne de McAllester du corollaire 3.2, la borne sur l'erreur réelle $\mathbf{R}_P(G_\rho)$ est explicitement liée à la somme de son estimateur $\mathbf{R}_S(G_\rho)$ et de $\mathrm{KL}(\rho\|\pi)$. D'un point de vue pratique, l'hyperparamètre C a l'avantage de pondérer l'importance de l'estimateur $\mathbf{R}_S(G_\rho)$ face à la complexité $\mathrm{KL}(\rho\|\pi)$. En effet, étant donné $C > 0$, optimiser la borne (3.3) revient à trouver la distribution ρ qui minimise : $Cm\mathbf{R}_S(G_\rho) + \mathrm{KL}(\rho\|\pi)$. Notons que les bornes Langford-Seeger du corollaire 3.1 et de Catoni du corollaire 3.3 coïncident en une valeur de C qui correspond à :

Proposition 3.2 ([Lacasse, 2010]) *Pour tout* $0 \leq \mathbf{R}_S(G_\rho) \leq \mathbf{R}_P(G_\rho) < 1$, *on a :*

$$\max_{C\geq 0} \left\{ -\ln\left(1 - \mathbf{R}_S(G_\rho)\left[1 - \exp\left(-C\right)\right]\right) - C\mathbf{R}_P(G_\rho) \right\} = \mathrm{kl}(\mathbf{R}_S(G_\rho), \mathbf{R}_P(G_\rho)).$$

Enfin, comme nous allons le voir dans la section 3.3, si ρ est une gaussienne isotrope sur l'espace des classifieurs linéaires, cette minimisation (sans information *a priori*) est étroitement liée à celle du problème associé aux SVM. L'algorithme PBGD qui en découle a été proposé par [Germain *et al.*, 2009a] et exploite ces résultats.

3.3 PBGD : Un algorithme de minimisation du théorème PAC-Bayes spécialisé aux classifieurs linéaires

Supposons maintenant que $\mathcal{H}$ est un ensemble de classifieurs linéaires $h(\mathbf{x}) = \mathrm{sign}\left(\langle \mathbf{v}, \mathbf{x}\rangle\right)$ avec $\mathbf{v} \in \mathbb{R}^d$ un vecteur de poids. En restreignant les distributions prior et posterior basées sur des gaussiennes, [Langford et Shawe-Taylor, 2002, Ambroladze *et al.*, 2006, Parrado-Hernández *et al.*, 2012] ont spécialisé la théorie PAC-Bayésienne pour borner l'erreur réelle de tout classifieur linéaire identifié par un vecteur de poids $\mathbf{w}$. Plus précisément, on considère un prior $\pi_{\mathbf{0}}$ et un posterior $\rho_{\mathbf{w}}$ définis comme une gaussienne sphérique de matrice de covariance égale à l'identité centrée respectivement sur les vecteurs $\mathbf{0}$ et $\mathbf{w}$. Plus formellement, pour tout $h = \mathrm{sign}\left(\langle \mathbf{v}, \mathbf{x}\rangle\right)$ issu de $\mathcal{H}$, on a :

$$\pi_{\mathbf{0}}(h) = \left(\tfrac{1}{\sqrt{2\Pi}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{v}\|^2\right),$$
$$\rho_{\mathbf{w}}(h) = \left(\tfrac{1}{\sqrt{2\Pi}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{w}\|^2\right).$$

L'erreur réelle du classifieur de Gibbs $G_{\rho_{\mathbf{w}}}(\cdot)$ sur un domaine P est alors donnée par :

$$\begin{aligned}\mathbf{R}_P(G_{\rho_{\mathbf{w}}}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathbf{I}(h \neq y) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \ell_{\mathbf{Erf}}\left(G_{\rho_{\mathbf{w}}}(\mathbf{x}), y\right),\end{aligned}$$

où :

$$\ell_{\mathbf{Erf}}\left(G_{\rho_{\mathbf{w}}}(\mathbf{x}), y\right) = \frac{1}{2}\left[1 - \mathbf{Erf}\left(\frac{1}{\sqrt{2}}\frac{y\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right)\right],$$

et $\mathbf{Erf}(a) = \frac{2}{\sqrt{\Pi}}\int_0^a \exp\left(-t^2\right) dt$ est la fonction d'erreur de Gauss. Dans cette situation, la KL-divergence entre $\rho_{\mathbf{w}}$ et $\pi_{\mathbf{0}}$ devient simplement :

$$\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}}) = \frac{1}{2}\|\mathbf{w}\|^2,$$

et le corollaire 3.1 se spécialise à :

Corollaire 3.4 ([Langford, 2005]) *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de classifieurs linéaires, soit S un échantillon de m éléments tirés i.i.d. selon P. Alors pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, pour tout vecteur posterior $\mathbf{w} \in \mathbb{R}^d$, on a :*

$$\mathrm{kl}\left(\mathbf{R}_S(G_{\rho_\mathbf{w}}) \| \mathbf{R}_P(G_{\rho_\mathbf{w}})\right) \leq \frac{1}{m}\left[\frac{1}{2}\|\mathbf{w}\|^2 + \ln\frac{m+1}{\delta}\right]. \tag{3.4}$$

Notons, que [Jin et Wang, 2012] ont étendu ce résultat en prenant en compte la dimension d de l'espace de description $X \in \mathbb{R}^d$. La borne obtenue est :

$$\mathrm{kl}\left(\mathbf{R}_S(G_{\rho_\mathbf{w}}) \| \mathbf{R}_P(G_{\rho_\mathbf{w}})\right) \leq \frac{1}{m}\left[\frac{d}{2}\ln\left(1+\frac{\|\mathbf{w}\|^2}{d}\right) + \ln\frac{m+1}{\delta}\right]. \tag{3.5}$$

Cette borne est monotone croissante en fonction de d. De plus, la borne (3.5) est plus précise que (3.4) lorsque d est fini. Notons que les deux bornes tendent a être équivalentes lorsque d tend vers $+\infty$.

À la manière du corollaire 3.4, la borne de Catoni du corollaire 3.3 se réécrit :

Corollaire 3.5 *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de classifieurs linéaires, soit S un échantillon de m éléments tirés i.i.d. selon P. Alors pour toute distribution prior π sur $\mathcal{H}$, pour tout réel $C > 0$ et pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, pour tout vecteur posterior $\mathbf{w} \in \mathbb{R}^d$, on a :*

$$\mathbf{R}_P(G_{\rho_\mathbf{w}}) \leq \frac{1}{1-e^{-C}}\left[C\mathbf{R}_S(G_{\rho_\mathbf{w}}) + \frac{\frac{1}{2}\|\mathbf{w}\|^2 + \ln\frac{1}{\delta}}{m}\right]. \tag{3.6}$$

En se basant sur cette spécialisation, [Germain *et al.*, 2009a] ont proposé d'optimiser les bornes (3.4) et (3.6) pour apprendre une distribution $\rho_\mathbf{w}$ performante. Nous présentons uniquement[7] l'approche visant à minimiser la borne (3.6) (à la "Catoni") permettant de contrôler le compromis erreur empirique/complexité.

Étant donnés un échantillon d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ dont les éléments sont *i.i.d* selon P et un hyperparamètre $C > 0$, la recherche du classifieur linéaire idéal s'exprime par le problème d'optimisation suivant :

$$\min_{\mathbf{w}} \; Cm\mathbf{R}_S(G_{\rho_\mathbf{w}}) + \mathrm{KL}(\rho_\mathbf{w}\|\pi_\mathbf{0}) \iff \min_{\mathbf{w}} \; C\sum_{i=1}^m \ell_{\mathbf{Erf}}\left(G_{\rho_\mathbf{w}}(\mathbf{x}_i), y_i\right) + \frac{1}{2}\|\mathbf{w}\|^2. \tag{3.7}$$

Concrètement, résoudre ce problème est équivalent à rechercher le meilleur compromis entre le risque empirique, mesuré via la fonction de perte $\ell_{\mathbf{Erf}}(\cdot,\cdot)$, et la complexité du classifieur linéaire appris, exprimée par le régularisateur $\|\mathbf{w}\|^2$. L'algorithme, nommé PBGD3, réalise une descente de gradient pour trouver le vecteur de poids optimal. Notons que l'astuce du noyau présentée en section 1.4.2 peut aisément être appliquée à PBGD3. Finalement, à la fonction de perte près, le problème (3.7) est très proche du problème d'optimisation des SVM (voir l'équation (1.7) pour rappel).

7. PBGD1 et PBGD2 ont été développés pour minimiser la borne du corollaire 3.4, c'est-à-dire sans paramètre de compromis. Pour plus de détails le lecteur peut se référer à l'article [Germain *et al.*, 2009a].

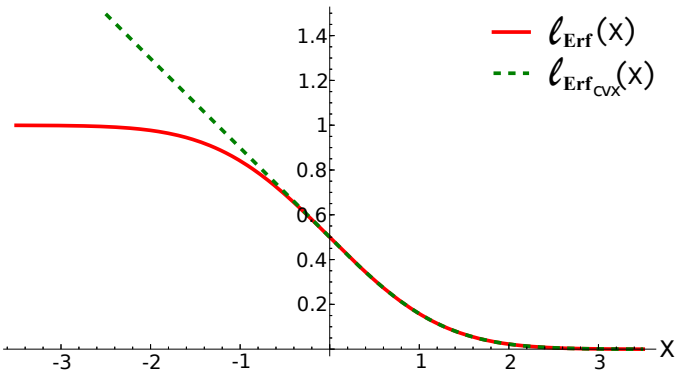


FIGURE 3.1 – *Comportement des fonctions* $\ell_{\mathbf{Erf}}(\cdot)$ *et* $\ell_{\mathbf{Erf}_{\text{cvx}}}(\cdot)$.

En pratique, la fonction objectif de PBGD3 n'est pas convexe. L'implémentation de la descente de gradient requiert donc de nombreux redémarrages. Cependant, dans ce manuscrit, nous remplacerons la fonction de perte $\ell_{\mathbf{Erf}}(\cdot,\cdot)$ par sa relaxation convexe (illustrée sur la figure 3.1) :

$$\ell_{\mathbf{Erf}_{\text{cvx}}}\left(G_{\rho_{\mathbf{w}}}(\mathbf{x}), y\right) = \begin{cases} \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \frac{y\langle \mathbf{w}, \mathbf{x}_i \rangle}{\|\mathbf{x}_i\|} & \text{si } \frac{y\langle \mathbf{w}, \mathbf{x}_i \rangle}{\|\mathbf{x}_i\|} \leq 0, \\ \ell_{\mathbf{Erf}}\left(G_{\rho_{\mathbf{w}}}(\mathbf{x}), y\right) & \text{sinon.} \end{cases}$$

En effet, il est empiriquement constaté que cette relaxation implique des performances similaires tout en étant plus rapide.

Malgré ses bons résultats empiriques, PBGD3 montre deux inconvénients. D'une part, il est spécifique aux classifieurs linéaires : la KL-divergence peut être difficile à calculer si les distributions ρ et π ne sont pas des Gaussiennes. D'autre part, il se focalise sur la minimisation de l'erreur du classifieur stochastique de Gibbs alors qu'en général nous nous intéressons au vote de majorité. Cependant, ces spécificités vont nous permettre de proposer, dans le chapitre 7, le premier algorithme PAC-Bayésien pour le problème de l'adaptation de domaine lorsque les données de test sont tirées *i.i.d* selon un domaine différent des données d'apprentissage.
Nous allons maintenant étudier un algorithme qui, quant à lui, optimise directement l'erreur du vote de majorité via la C-borne du théorème 3.1 sur un ensemble de votants quelconques (à valeurs réelles).

3.4 MinCq : Un algorithme de minimisation de l'erreur du vote de majorité

La théorie PAC-Bayésienne et, par la même occasion, PBGD3 se concentrent sur l'erreur du classifieur de Gibbs. Or, en pratique, notre objectif est plutôt d'apprendre un vote de majorité performant. C'est pourquoi, [Laviolette *et al.*, 2011a] se sont focalisés directement sur la minimisation de l'erreur du vote de majorité ρ-pondéré. Sachant que la C-borne de [Lacasse *et al.*, 2007], présentée dans le théorème 3.1, est un bon estimateur de cette erreur, l'optimiser apparaît être une bonne stratégie. Pour ce faire, les auteurs l'ont, tout d'abord, généralisée à un ensemble de votants à valeurs réelles,

puis ont démontré que sa minimisation empirique revient à résoudre un programme quadratique simple, justifiée par une borne PAC-Bayésienne sans terme de complexité $\mathrm{KL}(\rho\|\pi)$.

3.4.1 La C-borne : une majoration de l'erreur du vote de majorité sur un ensemble de votants réels

Nous supposons maintenant que $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$ est un ensemble de $2n$ votants à valeurs réelles : $\forall j \in \{1, \ldots, 2n\},\ h_j : X \mapsto \mathbb{R}$. Puisque nous sommes en classification binaire avec $Y = \{-1, +1\}$, la définition du vote de majorité $B_\rho(\cdot)$ reste identique et celle du classifieur de Gibbs est légèrement modifiée : on ne renvoie plus directement $h(\mathbf{x})$, mais $\mathrm{sign}\,(h(\mathbf{x}))$. La généralisation de la C-borne à des votants réels repose simplement sur le lien qu'il existe entre l'erreur $\mathbf{R}_P(B_\rho)$ et la confiance de $B_\rho(\cdot)$ en son étiquetage, modélisée par la notion suivante de ρ-marge. Cette relation permet de définir la C-borne en fonction des premier et second moments statistiques de la ρ-marge [Laviolette *et al.*, 2011a].

Définition 3.2 (ρ-marge) *La ρ-marge de $B_\rho(\cdot)$ mesurée sur un exemple $(\mathbf{x}, y)$ est :*

$$\mathcal{M}^\rho(\mathbf{x}, y) = y \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}).$$

Soit P un domaine sur $X \times Y$. Le premier moment $\mathcal{M}^\rho_P$ de la ρ-marge sur P est :

$$\begin{aligned}\mathcal{M}^\rho_P &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathcal{M}^\rho(\mathbf{x}, y)\\ &= \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} yh(\mathbf{x}),\end{aligned}$$

alors que le second moment $\mathcal{M}^{\rho^2}_P$ de la ρ-marge sur P est défini ainsi :

$$\begin{aligned}\mathcal{M}^{\rho^2}_P &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y)\right)^2\\ &= \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} h(\mathbf{x})h'(\mathbf{x}).\end{aligned}$$

Les premier et second moments empiriques $\mathcal{M}^\rho_S$ et $\mathcal{M}^{\rho^2}_S$ estimés sur un échantillon $S \sim (P)^m$ sont :

$$\mathcal{M}^\rho_S = \frac{1}{m}\sum_{i=1}^m \mathcal{M}^\rho(\mathbf{x}_i, y_i), \quad \textit{et} \quad \mathcal{M}^{\rho^2}_S = \frac{1}{m}\sum_{i=1}^m \left(\mathcal{M}^\rho(\mathbf{x}_i, y_i)\right)^2.$$

D'après cette définition, $B_\rho(\cdot)$ classe correctement un exemple $(\mathbf{x}, y)$ lorsque sa ρ-marge est strictement positive :

$$\mathbf{R}_P(B_\rho) = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y) \leq 0\right). \tag{3.8}$$

Cette égalité permet de dériver simplement la version suivante de la C-borne en appliquant l'inégalité de Cantelli-Chebitchev [8] sur la variable aléatoire $\mathcal{M}^\rho(\mathbf{x}, y)$. Notons que dans le chapitre 5, nous généraliserons ce résultat au cadre multiclasse.

8. L'inégalité de Cantelli-Chebitchev est énoncée dans le théorème A.2 en annexe A.

Théorème 3.3 (C-borne de [Laviolette *et al.*, 2011a]) *Pour toute distribution ρ sur un ensemble de votants à valeurs réelles $\mathcal{H}$ et pour tout domaine P sur $X \times Y$, si le premier moment est strictement positif, $\mathcal{M}_P^\rho > 0$, alors on a :*

$$\mathbf{R}_P(B_\rho) \leq C_P^\rho,$$

avec :

$$C_P^\rho = \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x},y)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\, \left(\mathcal{M}^\rho(\mathbf{x},y)\right)^2}.$$
$$= 1 - \frac{\left(\mathcal{M}_P^\rho\right)^2}{\mathcal{M}_P^{\rho^2}}.$$

Nous notons $C_S^\rho = 1 - \frac{\left(\mathcal{M}_S^\rho\right)^2}{\mathcal{M}_S^{\rho^2}}$ son estimation sur l'échantillon $S \sim (P)^m$.

Démonstration. Voir en annexe B.3. □

Cette version généralise bien le premier résultat du théorème 3.1, puisque trivialement si les votants sont binaires, c'est-à-dire si pour tout h issu de $\mathcal{H}$, $h : X \mapsto \{-1, +1\}$, on a :

$$\mathcal{M}_P^\rho = 1 - 2\mathbf{R}_P(G_\rho)\ , \quad \text{et} \quad \mathcal{M}_P^{\rho^2} = 1 - 2\mathbf{R}_P(G_\rho, G_\rho).$$

La C-borne est en fait connue pour être un estimateur précis de l'erreur du vote majorité [Lacasse *et al.*, 2007], sa minimisation est donc une solution naturelle pour apprendre une distribution ρ menant à un vote de majorité ρ-pondéré $B_\rho(\cdot)$ d'erreur réelle faible. Pour justifier cette stratégie, [Laviolette *et al.*, 2011a] ont dérivé une borne en généralisation PAC-Bayésienne sur C_P^ρ sans terme de complexité $\mathrm{KL}(\rho\|\pi)$, portant sur la notion de marge (et non plus celle d'erreur).

3.4.2 De la C-borne à l'algorithme MinCq

Pour s'affranchir du terme $\mathrm{KL}(\rho\|\pi)$, on se focalise sur des distributions ρ quasi-uniformes définies sur un ensemble auto-complémenté de $2n$ votants $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$. Plus formellement, pour tout $j \in \{1, \ldots, n\}$, on suppose :

$$h_{j+1} = -h_j \qquad \text{(auto-complémentation)}, \tag{3.9}$$
$$\rho(h_j) + \rho(h_{j+n}) = \frac{1}{n} \qquad \text{(quasi-uniformité)}. \tag{3.10}$$

Ces hypothèses permettent de caractériser les situations pour lesquelles on suppose le même *a priori* sur chaque couple de votants (h_j, h_{j+n}) (le prior π est non informatif puisqu'il affecte la même importance à chaque couple) et n'impliquent donc pas une trop forte restriction. De plus, les distributions quasi-uniformes sur un ensemble auto-complémenté révèlent deux avantages.

- D'une part, nous verrons que ces hypothèses nous permettrons de s'affranchir de la KL-divergence (ce terme, parfois difficile à optimiser, peut amener à une mauvaise régularisation).
- D'autre part, cette contrainte peut être vue comme une régularisation donnant le même *a priori* à chaque couple de votants et fournit une solution concrète et naturelle au sur-apprentissage.

D'après le théorème 3.3, la borne en généralisation est obtenue en prenant la minoration (respectivement la majoration) de $\mathcal{M}_P^\rho$ et la majoration (respectivement la minoration) de $\mathcal{M}_P^{\rho^2}$ du théorème suivant.

Théorème 3.4 ([Laviolette *et al.*, 2011a]) *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble auto-complémenté de votants réels tel que : $\forall h \in \mathcal{H},\ |h_j(\mathbf{x}) \leq B$.Soit $m \geq 8$, soit S un échantillon de m éléments i.i.d. selon P. Alors pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0, 1]$, avec une probabilité d'au moins $1 - \delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, pour toute distribution posterior ρ sur $\mathcal{H}$, on a :*

$$\left|\mathcal{M}_P^\rho - \mathcal{M}_S^\rho\right| \leq \frac{2B\sqrt{\ln \frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}}, \quad \textit{et} \quad \left|\mathcal{M}_P^{\rho^2} - \mathcal{M}_S^{\rho^2}\right| \leq \frac{2B^2\sqrt{\ln \frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}}.$$

Principe de la démonstration (la preuve complète est disponible dans [Laviolette et al., 2011b]). On suit le même principe que la preuve du corollaire 3.2 en posant : $\mathcal{D}(a, b) = \frac{1}{2B^2}(a - b)^2$. Les principales différences sont :

- On considère $a = \mathcal{M}_S^\rho$ et $b = \mathcal{M}_P^\rho$ au lieu de $\mathbf{R}_S(G_\rho)$ et $\mathbf{R}_P(G_\rho)$.
- $\mathcal{H}$ est un ensemble auto-complémenté de votants réels tel que $\forall h \in \mathcal{H},\ |h(\mathbf{x})| \leq B$. Une distribution sur $\mathcal{H}$ est quasi-uniforme si pour tout h issu de $\mathcal{H}$, on a :

$$\begin{aligned} \rho(h) + \rho(-h) &= \pi(h) + \pi(-h) \\ &= \frac{1}{n}. \end{aligned} \tag{3.11}$$

De plus, en posant :

$$\mathcal{M}_P^h = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} yh(\mathbf{x}), \quad \text{et} \quad \mathcal{M}_S^h = \frac{1}{m}\sum_{(\mathbf{x},y)\in S} yh(\mathbf{x}),$$

cela implique :

$$\mathcal{M}_P^{-h} = -\mathcal{M}_P^h.$$

Ainsi :

$$\begin{aligned} \left(\mathcal{M}_S^{-h} - \mathcal{M}_P^{-h}\right)^2 &= \left(-\mathcal{M}_S^h - \left(-\mathcal{M}_P^h\right)\right)^2 \\ &= \left(\mathcal{M}_S^h - \mathcal{M}_P^h\right)^2. \end{aligned}$$

Finalement, pour passer de la distribution π à ρ, on regroupe les termes par couples $(h, -h)$, puis il suffit de remplacer $\pi(h) + \pi(-h)$ par $\rho(h) + \rho(-h)$ en utilisant l'équation 3.11. Cette équation implique alors $\mathrm{KL}(\rho\|\pi) = 0$.

- On applique le lemme de Maurer, énoncé dans le lemme A.1 en annexe A, pour majorer la partie droite de la borne par $2\sqrt{m}$, pour $m \geq 8$.

□

Les hypothèses de quasi-uniformité et d'auto-complémentation n'impliquent pas de perte d'expressivité. Ceci fournit à cette approche des propriétés extrêmement élégantes. Cet aspect est formalisé par la proposition suivante.

Proposition 3.3 ([Laviolette *et al.*, 2011a]) *Pour tout $\mu \in (0,1]$ et pour toute distribution ρ sur $\mathcal{H}$ associée à une ρ-marge empirique $\mathcal{M}_S^\rho \geq \mu$, il existe une distribution quasi-uniforme ρ' sur $\mathcal{H}$ de ρ'-marge empirique égale à μ, telle que ρ et ρ' induisent le même vote de majorité pondéré et la même valeur empirique de la C-borne, c'est-à-dire :*

$$\mathcal{M}_S^{\rho'} = \mu, \qquad B_{\rho'} = B_\rho, \qquad C_S^\rho = C_S^\rho, \quad \textit{et} \quad C_P^{\rho'} = C_P^\rho.$$

En se basant sur les théorèmes 3.3, 3.4 et sur la proposition 3.3, ainsi que pour éviter les instabilités numériques dues à la forme indéterminée "$\frac{0}{0}$", [Laviolette *et al.*, 2011a] proposent de minimiser la C-borne empirique sous la contrainte $\mathcal{M}_S^\rho = \mu$, ou autrement dit en fixant la valeur de la marge à atteindre.
Soit un échantillon d'apprentissage S constitué de m éléments tirés *i.i.d.* selon P, soit $\mathcal{H}$ un ensemble auto-complémenté, soit $\mu > 0$. Grâce à la quasi-uniformité, trouver la distribution posterior ρ sur $\mathcal{H}$ minimisant la C-borne est équivalent à trouver le vecteur de poids $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)^\top$ (où $\rho_i = \rho(h_i)$) qui minimise le programme quadratique simple suivant en ne mettant uniquement en jeu que les n premiers votants de $\mathcal{H}$.

$$\begin{cases} \min\limits_{\boldsymbol{\rho}} & \boldsymbol{\rho}^\top \mathbf{M}_S \boldsymbol{\rho} - \mathbf{A}_S{}^\top \boldsymbol{\rho}, \\ \text{s.c.} & \mathbf{m}_S{}^\top \boldsymbol{\rho} = \dfrac{\mu}{2} + \dfrac{1}{2nm} \displaystyle\sum_{j=1}^{n} \sum_{i=1}^{m} y_i h_j(\mathbf{x}_i), \\ & \forall j \in \{1, \ldots, n\}, \quad 0 \leq \rho_j \leq \dfrac{1}{n}, \end{cases} \tag{3.12}$$

où $\mathbf{M}_S$ est la matrice $n \times n$ dont les éléments d'indices (j, j') sont définis par :

$$\frac{1}{m} \sum_{i=1}^{m} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i),$$

et :

$$\mathbf{m}_S = \left(\frac{1}{m} \sum_{i=1}^{m} y_i h_1(\mathbf{x}_i), \ldots, \frac{1}{m} \sum_{i=1}^{m} y_i h_n(\mathbf{x}_i) \right)^\top,$$

$$\mathbf{A}_S = \left(\frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} h_1(\mathbf{x}_i) h_j(\mathbf{x}_i), \ldots, \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} h_n(\mathbf{x}_i) h_j(\mathbf{x}_i) \right)^\top.$$

Résoudre ce problème revient à minimiser le dénominateur $\mathcal{M}_S^{\rho^2}$, c'est-à-dire le second moment de la ρ-marge (fonction objectif) — sous les contraintes $\mathcal{M}_S^\rho = \mu$ (première contrainte), c'est-à-dire le premier moment est fixé, et ρ est quasi-uniforme (seconde contrainte). Finalement le vote de majorité ρ-pondéré appris est :

$$B_\rho(\mathbf{x}) = \operatorname{sign}\left[\sum_{j=1}^{n} \left(2\rho_j - \frac{1}{n} \right) h_j(\mathbf{x}) \right].$$

Version	$2\mathcal{D}\left(\mathbf{R}_S(G_\rho), \mathbf{R}_P(G_\rho)\right)$	$\mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} e^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}$	spécificités
McAllester	$\left(\mathbf{R}_S(G_\rho) - \mathbf{R}_P(G_\rho)\right)\right)^2$	$\leq\ m+1$	relation linéaire
Seeger	$\mathrm{kl}\left(\mathbf{R}_S(G_\rho), \mathbf{R}_P(G_\rho)\right)$	$\leq\ m+1$	plus précise que celle de McAllester
Catoni	$\ln \frac{1}{1-\mathbf{R}_P(G_\rho)\,[1-\exp(-C)]} - C\mathbf{R}_S(G_\rho)$	$=\ 1$	paramétrée, $C>0$

TABLE 3.1 – *Résumé des spécificités des trois versions classiques du théorème PAC-Bayes.*

Cet algorithme a montré de bons résultats sur des votes de stumps et de noyaux gaussiens avec un prior non informatif. Cependant, la borne PAC-Bayésienne du théorème 3.4 n'est pas valide dans ce cas, car les votants sont définis à partir des exemples d'apprentissage (on parle alors de schéma de compression [Floyd et Warmuth, 1995]).

En ayant pour objectif de combiner différents classifieurs (ou régresseurs), nous allons proposer une amélioration de MinCq de deux manières dans le chapitre 4. Tout d'abord nous généraliserons la contrainte de quasi-uniformité à une contrainte nous permettant de modéliser la connaissance *a priori* sur la pertinence des votants. Puis nous étendrons le théorème 3.4 aux schémas de compression PAC-Bayésiens [Graepel *et al.*, 2005] (lorsque les votants dépendent des données d'apprentissage). Pour ce faire, nous suivrons l'intuition donnée dans [Laviolette *et al.*, 2011a] en faisant appel aux techniques de [Laviolette et Marchand, 2007].

3.5 SYNTHÈSE

Nous avons introduit les bases de la théorie PAC-Bayésienne qui offre des garanties en généralisation pour des votes de majorité sur un ensemble d'hypothèses $\mathcal{H}$. Contrairement aux approches classiques présentées en section 1.3, cette théorie à l'avantage de directement considérer un moyennage sur $\mathcal{H}$ selon une distribution ρ plutôt que de se focaliser sur chacune des hypothèses. Comme illustré dans la table 3.1, l'aspect générique du théorème PAC-Bayes permet de considérer différentes mesures de comparaison entre valeur réelle et estimation du risque, en fonction du problème qui nous intéresse. Finalement, nous pensons que cette théorie PAC-Bayésienne offre un cadre théorique naturel et élégant pour le problème de l'apprentissage de vote de majorité sur des classifieurs qui est au cœur de ce mémoire. Tout d'abord, dans le chapitre 4 suivant, nous étendons MinCq, d'une part, pour qu'il considère un *a priori* informatif, d'autre part, à des votants dépendants des données. Nous en montrons aussi son intérêt pour combiner différents types de fonctions. Ensuite, dans le chapitre 5 nous nous concentrerons sur la classification multiclasse. Nous démontrerons la première borne PAC-Bayésienne sur la matrice de confusion, puis nous généraliserons la C-borne au multiclasse. Enfin, dans le chapitre 7, nous proposerons la première analyse PAC-Bayésienne pour le problème de l'adaptation de domaine que nous avons présenté dans le chapitre précédent.

Deuxième partie

Contributions en apprentissage supervisé

Vote de majorité contraint et classification binaire

4

Après avoir introduit dans la première partie du mémoire les notions qui nous seront utiles tout au long de cette thèse, nous présentons dans ce quatrième chapitre, notre première contribution : une généralisation de l'algorithme MinCq [Laviolette *et al.*, 2011a] pour contrer les limitations énoncées en section 3.4. Nous rappelons que cet algorithme élégant trouve sa source dans la théorie PAC-Bayésienne du chapitre précédent et est, en ce sens, un algorithme d'apprentissage de vote de majorité pondéré. L'objectif est alors de construire un vote final plus performant et plus robuste que les votants individuels. Alors que le choix des poids pertinents sur l'ensemble des votants $\mathcal{H}$ est parfois une tâche complexe, MinCq les optimise en minimisant la C-borne du théorème 3.3 — sur l'erreur du vote de majorité — mettant en jeu les deux premiers moments statistiques de la marge du vote. La sortie de l'algorithme MinCq est une distribution posterior ρ sur $\mathcal{H}$ pondérant les votants sur

un ensemble auto-complémenté [1].

Notre généralisation de MinCq est présentée en section 4.1. Tout d'abord, alors que MinCq utilise un prior non informatif en affectant la même importance à toutes les paires votants/opposés (hypothèse de quasi-uniformité), nous reformulons le problème pour contraindre la distribution posterior à être $\boldsymbol{\pi}$-alignée, où $\boldsymbol{\pi}$ est un vecteur adapté à MinCq et qui modélise un *a priori* sur l'importance de chaque paire. $\boldsymbol{\pi}$ permet ainsi d'incorporer la connaissance dont on peut disposer en amont sur la performance de chaque votant. Nous montrons que toute distribution sur $\mathcal{H}$ peut être exprimée comme une distribution $\boldsymbol{\pi}$-alignée et que ce nouveau problème, appelé P-MinCq, se formule lui aussi comme un programme quadratique. Ensuite, nous étendons la borne en généralisation PAC-Bayésienne de MinCq énoncée dans le théorème 3.4 du chapitre précédent aux schémas de compression, c'est-à-dire lorsque les votants sont définis à partir des données d'apprentissage. Puis, dans la section 4.2 nous nous attaquons aux inconvénients de la classification par k-PPV en proposant une instanciation de P-MinCq visant à combiner plusieurs classifieurs k-PPV. Enfin, en section 4.3 nous spécialisons l'approche à la tâche de fusion de classifieurs [2] dans un contexte de recherche d'information multimédia, lorsque les votants à combiner sont appris à partir de différentes descriptions des données.

Les travaux présentés dans ce chapitre ont donné lieu à une publication à la conférence CAp 2013 [Bellet *et al.*, 2013a] et ont été soumis dans un journal.

4.1 Extension de MinCq à P-MinCq

L'extension de MinCq que nous proposons se présente selon deux axes. Premièrement, dans la section 4.1.1 nous étendons la contrainte de quasi-uniformité (non informative, voir équation (3.10)) à une contrainte plus générale sur les distributions prior π et posterior ρ sur $\mathcal{H}$: étant donnée une contrainte $\boldsymbol{\pi}$ définie en amont de la phase d'apprentissage, nous nous concentrons sur des distributions $\boldsymbol{\pi}$-alignées. En prenant en compte dans l'algorithme la connaissance *a priori* portée par $\boldsymbol{\pi}$, nous obtenons le problème quadratique P-MinCq. Deuxièmement, dans la section 4.1.3, nous justifions de l'utilisation MinCq (et P-MinCq) lorsque les votants dépendent des données d'apprentissage.

4.1.1 D'une contrainte de quasi-uniformité à une contrainte de π-alignement

Similairement à MinCq, nous définissons $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$ comme étant un ensemble de votants auto-complémentés (c'est-à-dire : $\forall j \in \{1, \ldots, n\},\ h_{j+n} = -h_j$). Plutôt que de contraindre la distribution ρ sur $\mathcal{H}$ à être quasi-uniforme (c'est-à-dire : $\forall j \in \{1, \ldots, n\},\ \rho(h_j) + \rho(h_{j+n}) = \frac{1}{n}$), nous généralisons cette approche à toutes les distributions $\boldsymbol{\pi}$-alignées. Autrement dit, étant donné un vecteur $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ tel

1. Nous rappelons qu'un ensemble $\mathcal{H}$ est auto-complémenté lorsque pour tout votant h issu de $\mathcal{H}$, son opposé $-h$ est aussi un élément de $\mathcal{H}$.

2. En apprentissage automatique, on parle aussi d'apprentissage multimodal ou multivue.

que $\sum_{j=1}^{n} \pi_j = 1$, on veut :

$$\begin{aligned}\forall j \in \{1,\dots,n\},\ \rho(h_j) + \rho(h_{j+n}) &= \pi(h_j) + \pi(h_{j+n}) \\ &= \pi_j.\end{aligned}$$

C'est donc $\boldsymbol{\pi}$ qui va jouer le rôle d'une connaissance sur les votants. Remarquons que la contrainte de quasi-uniformité est un cas particulier du $\boldsymbol{\pi}$-alignement pour laquelle on a : $\forall j \in \{1,\dots,n\},\ \pi_j = \frac{1}{n}$. Par la suite, nous montrons qu'une contrainte de $\boldsymbol{\pi}$-alignement ne restreint pas l'ensemble des votes de majorité possibles. Puis, nous introduisons notre extension de MinCq, appelée P-MinCq, optimisant la C-borne du théorème 3.3 dans ce contexte.

Expressivité des distributions $\boldsymbol{\pi}$-alignées

Nous rappelons que dans [Laviolette *et al.*, 2011a], les auteurs ont montré que la contrainte de quasi-uniformité ne restreignait pas l'ensemble des votes de majorité possibles (voir section 3.4.2). Nous généralisons cette preuve à toutes les distributions $\boldsymbol{\pi}$-alignées sur un ensemble de votants auto-complémenté $\mathcal{H}$ et démontrons qu'une telle contrainte ne restreint pas les résultats que peut renvoyer un algorithme de minimisation de la C-borne.

Proposition 4.1 *Pour toute distribution ρ sur l'ensemble $\mathcal{H}$, il existe une distribution $\boldsymbol{\pi}$-alignée ρ' sur l'ensemble $\mathcal{H}$ auto-complémenté pour laquelle le vote de majorité $B_{\rho'}$ est identique à B_ρ et possédant la même C-borne (empirique et réelle) que ρ'.*

Démonstration. En annexe C.1. □

Puisque le $\boldsymbol{\pi}$-alignement est une généralisation de la quasi-uniformité, le principe de la proposition 3.3 du chapitre 3 s'applique trivialement : sous la contrainte $\mathcal{M}_S^\rho = \mu > 0$, la C-borne peut être optimisée en minimisant le second moment $\mathcal{M}_S^{\rho^2}$ de la ρ-marge. En effet, si ρ est une distribution $\boldsymbol{\pi}$-alignée sur $\mathcal{H}$ telle que $\mathcal{M}_S^\rho \geq \mu$ et que l'on définit ρ' par :

$$\forall j \in \{1\dots,n\},\ \rho'(h_j) = \frac{\mu}{\mathcal{M}_S^\rho}\,\rho(h_j) + \left(1 - \frac{\mu}{\mathcal{M}_S^\rho}\right)\frac{\pi_j}{2},$$

alors on a :

$$\begin{aligned}\mathop{\mathbf{E}}_{h\sim\rho'} h(\mathbf{x}) &= \sum_{j=1}^{2n} \rho'(h_j) h_j(\mathbf{x}) \\ &= \sum_{j=1}^{n} \left[\rho'(h_j) - \rho'(h_{j+n})\right] h_j(\mathbf{x}) \\ &= \sum_{=1}^{n} \frac{\mu}{\mathcal{M}_S^\rho} \left[\rho(h_j) - \rho(h_{j+n})\right] h_j(\mathbf{x}) \\ &= \frac{\mu}{\mathcal{M}_S^\rho} \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}).\end{aligned}$$

Le programme quadratique P-MinCq présenté dans la section suivante permet cette minimisation.

4.1.2 P-MinCq : un programme quadratique de minimisation de la C-borne

La démonstration des bornes en généralisation du théorème 3.4 reste valide dans le cas de distribution $\boldsymbol{\pi}$-alignées sur un ensemble de votants indépendants des données d'apprentissage. En effet, la preuve utilise uniquement le fait que pour tout $j \in \{1,\ldots,n\}$, on a : $\rho(h_j) + \rho(-h_j) = \pi(h_j) + \pi(-h_j)$ (voir l'équation (3.11) dans le chapitre 3). Ceci correspond en fait à l'hypothèse de $\boldsymbol{\pi}$-alignement lorsque l'on pose $\pi_j = \pi(h_j) + \pi(-h_j)$. Nous pouvons donc généraliser l'algorithme MinCq à P-MinCq, décrit dans le problème (4.1). Soit un échantillon d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ constitué de m éléments tirés *i.i.d.* selon (P), $\mathcal{H}$ un ensemble auto-complémenté de votants, une marge $\mu > 0$ et un vecteur *a priori* $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_n)^\top$. Similairement à MinCq, grâce à la contrainte de $\boldsymbol{\pi}$-alignement, seuls les n premiers votants de $\mathcal{H}$ interviennent dans la résolution du programme quadratique. Trouver la distribution posterior ρ sur $\mathcal{H}$ minimisant la C-borne est alors équivalent à trouver le vecteur de poids $\boldsymbol{\rho} = (\rho_1,\ldots,\rho_n)^\top$ (où $\rho_j = \rho(h_j)$) qui minimise le programme quadratique simple suivant :

$$\begin{cases} \min\limits_{\boldsymbol{\rho}} & (\boldsymbol{\rho}-\boldsymbol{\pi})^\top \mathbf{M}_S \boldsymbol{\rho}, \\ \text{s.c.} & \mathbf{m}_S{}^\top (2\boldsymbol{\rho}-\boldsymbol{\pi}) = \mu, \\ & \forall j \in \{1,\ldots,n\}, \quad 0 \le \rho_j \le \pi_j, \end{cases} \tag{4.1}$$

où $\mathbf{M}_S$ est la matrice $n \times n$ dont les éléments d'indices $(j,j') \in \{1,\ldots,n\}^2$ sont définis par :

$$\frac{1}{m}\sum_{i=1}^m h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i)$$

et :

$$\mathbf{m}_S = \left(\frac{1}{m}\sum_{i=1}^m y_i h_1(\mathbf{x}_i), \ldots, \frac{1}{m}\sum_{i=1}^m y_i h_n(\mathbf{x}_i)\right)^\top.$$

Preuve du problème (4.1).

La fonction objectif. Nous montrons comment obtenir la fonction objectif à partir de la définition du second moment $\mathcal{M}_S^{\rho^2}$ de la ρ-marge mesurée sur l'échantillon S.

$$\begin{aligned}
\mathcal{M}_S^{\rho^2} &= \frac{1}{m} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \sum_{i=1}^m h(\mathbf{x}_i) h'(\mathbf{x}_i) \\
&= \frac{1}{m}\sum_{j=1}^{2n}\sum_{j'=1}^{2n}\sum_{i=1}^m \rho_j \rho_{j'} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i) \\
&= \frac{1}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^m \Big[\rho_j \rho_{j'} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i) + \rho_{j+n}\rho_{j'} h_{j+n}(\mathbf{x}_i) h_{j'}(\mathbf{x}_i) \\
&\qquad\qquad + \rho_j \rho_{j'+n} h_j(\mathbf{x}_i) h_{j'+n}(\mathbf{x}_i) + \rho_{j+n}\rho_{j'+n} h_{j+n}(\mathbf{x}_i) h_{j'+n}(\mathbf{x}_i)\Big] \\
&= \frac{1}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^m \Big[\rho_j \rho_{j'} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i) - \rho_{j+n}\rho_{j'} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i) \\
&\qquad\qquad - \rho_j \rho_{j'+n} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i) + \rho_{j+n}\rho_{j'+n} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i)\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{car } h_{j+n} = -h_j)
\end{aligned}$$

$$
\begin{aligned}
\mathcal{M}_S^{\rho^2} =& \frac{1}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^{m} h_j(\mathbf{x}_i)h_{j'}(\mathbf{x}_i)[\rho_j\rho_{j'} - (\pi_j-\rho_j)\rho_{j'} - \rho_j(\pi_{j'}-\rho_{j'}) + (\pi_j-\rho_j)(\pi_{j'}-\rho_{j'})] \\
=& \frac{1}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^{m} h_j(\mathbf{x}_i)h_{j'}(\mathbf{x}_i)[4\rho_j\rho_{j'} - 2\pi_j\rho_{j'} - 2\pi_{j'}\rho_j + \pi_j\pi_{j'}] \\
=& \frac{4}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^{m} \rho_j h_j(\mathbf{x}_i)h_{j'}(\mathbf{x}_i)\rho_{j'} - \frac{4}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^{m} \pi_j h_j(\mathbf{x}_i)h_{j'}(\mathbf{x}_i)\rho_{j'} \\
& \qquad + \frac{1}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^{m} \pi_j\pi_{j'} h_j(\mathbf{x}_i)h_{j'}(\mathbf{x}_i) \\
=& 4[(\boldsymbol{\rho}-\boldsymbol{\pi})^\top \mathbf{M}_S\boldsymbol{\rho}] + c_1,
\end{aligned}
$$

où $c_1 = \frac{1}{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\sum_{i=1}^{m} \pi_j\pi_{j'}h_j(\mathbf{x}_i)h_{j'}(\mathbf{x}_i)$ et la valeur multiplicative 4 sont considérés comme des constantes quelle que soit la distribution ρ considérée. On obtient ainsi la fonction objectif cherchée.

La contrainte sur la marge. Nous montrons maintenant comment obtenir la première contrainte à partir du premier moment $\mathcal{M}_S^\rho$ estimé sur S.

$$
\begin{aligned}
\mathcal{M}_S^\rho &= \frac{1}{m}\mathop{\mathbf{E}}_{h\sim\rho}\sum_{i=1}^{m} y_i h(\mathbf{x}_i) \\
&= \frac{1}{m}\sum_{j=1}^{2n}\sum_{i=1}^{m} \rho_j y_i h_j(\mathbf{x}_i) \\
&= \frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m} (\rho_j - \rho_{j+n}) y_i h_j(\mathbf{x}_i) \\
&= \frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m} (2\rho_j - \pi_j) y_i h_j(\mathbf{x}_i) \\
&= \mathbf{m}_S{}^\top(2\boldsymbol{\rho}-\boldsymbol{\pi}),
\end{aligned}
$$

où $\mathbf{m}_S{}^\top = \left(\frac{1}{m}\sum_{i=1}^{m} y_i h_1(\mathbf{x}_i), \ldots, \frac{1}{m}\sum_{i=1}^{m} y_i h_n(\mathbf{x}_i)\right)^\top$.
En remplaçant $\mathcal{M}_S^\rho$ par μ, on retrouve la première contrainte. □

La fonction objectif minimise le second moment de la ρ-marge alors que la première contrainte force la marge à être égale à μ. On peut noter que la partie gauche de cette contrainte est une moyenne pondérée (dont les poids sont égaux à $2\rho_j - \pi_j$) des ρ-marges des votants pris individuellement. Étant donnée $\boldsymbol{\pi}$, la dernière contrainte permet de ne considérer que des distributions $\boldsymbol{\pi}$-alignées. Finalement, le vote de majorité ρ-pondéré appris par P-MinCq est :

$$
B_\rho(\mathbf{x}) = \operatorname{sign}\left[\sum_{i=1}^{n} \left(2\rho_j - \pi_j\right) h_j(\mathbf{x})\right].
$$

Concernant les capacités en généralisation de P-MinCq, le théorème 3.4 du chapitre 3 n'est pas valide lorsque les votants dépendent des données d'apprentissage, autrement dit lorsque l'on considère le cadre des schémas de compression.

Cependant, les auteurs de [Laviolette *et al.*, 2011a] ont argumenté que l'on pouvait étendre les résultats à ce contexte particulier en utilisant des techniques issues de [Laviolette et Marchand, 2007]. Nous proposons donc dans la section suivante de suivre cette intuition pour généraliser le théorème 3.4 aux schémas de compression.

4.1.3 Borne en généralisation pour les schémas de compression

Nous dérivons maintenant une preuve de consistance pour MinCq et son extension π-alignée P-MinCq lorsque les votants sont définis à partir d'exemples de l'échantillon d'apprentissage, ce qui correspond à l'appellation "schéma de compression".

Cadre général d'un schéma de compression

Un schéma de compression [Floyd et Warmuth, 1995] est un algorithme d'apprentissage $\mathcal{A}$ travaillant sur un ensemble de classifieurs dépendant des données. Un classifieur est alors représenté par deux éléments :

(i) une séquence d'exemples, appelée la séquence de compression ;

(ii) un message représentant l'information supplémentaire utilisée permettant d'obtenir le classifieur à partir de la séquence de compression.

On définit ensuite une fonction de reconstruction capable de renvoyer un classifieur à partir d'une séquence de compression et d'un message.
Plus formellement, un algorithme $\mathcal{A}$ est un schéma de compression s'il vérifie la définition suivante.

Définition 4.1 *Soit $S \in (X \times Y)^m$ un échantillon d'apprentissage de taille m. On définit $\mathbf{I}_m$ l'ensemble de tous les vecteurs d'indices possibles :*

$$\mathbf{I}_m = \bigcup_{i=1}^{m} \left\{ (j_1, \ldots, j_i) \in \{1, \ldots, m\}^i \right\}.$$

Étant donnés une famille d'hypothèse $\mathcal{H}^S$ de X vers Y et un vecteur d'indices $\mathbf{i} \in \mathbf{I}_m$, on définit la séquence de compression $S_{\mathbf{i}}$ comme étant la sous-séquence indicée par $\mathbf{i}$:

$$S_{\mathbf{i}} = \left((\mathbf{x}_{j_1}, y_{j_1}), \ldots, (\mathbf{x}_{j_i}, y_{j_i}) \right).$$

Un algorithme $\mathcal{A} : (X \times Y)^{(\infty)} \mapsto \mathcal{H}^S$ est un schéma de compression si et seulement s'il existe un triplet $(\mathcal{C}, \mathcal{R}, \omega)$ tel que pour tout échantillon d'apprentissage S, on ait :

$$\mathcal{A}(S) = \mathcal{R}\left(S_{\mathcal{C}(S)}, \omega \right),$$

où $\mathcal{C} : (X \times Y)^{(\infty)} \mapsto \bigcup_{m=1}^{\infty} \mathbf{I}_m$ est la fonction de compression, $\mathcal{R} : (X \times Y)^{(\infty)} \times \Omega_{S_{\mathcal{C}(S)}} \mapsto \mathcal{H}^S$ est la fonction de reconstruction et ω est un message choisi dans l'ensemble $\Omega_{S_{\mathcal{C}(S)}}$ (défini a priori) de tous les messages qui peuvent être fournis avec la séquence de compression $S_{\mathcal{C}(S)}$ pour permettre la reconstruction du classifieur.

En d'autres termes, un schéma de compression est une fonction de reconstruction $\mathcal{R}(\cdot,\cdot)$ associant une séquence de compression $S_{\mathcal{C}(S)} = S_{\mathbf{i}}$ à un ensemble $\mathcal{H}^S$ de fonctions $h^{\omega}_{S_{\mathbf{i}}}$ telles que $\mathcal{A}(S) = \mathcal{R}\left(S_{\mathbf{i}}, \omega\right) = h^{\omega}_{S_{\mathbf{i}}}$.
Par exemple, les classifieurs de type plus proches voisins sont des classifieurs reconstructibles uniquement à partir d'une séquence de compression encodant les PPV (voir [Floyd et Warmuth, 1995, Graepel *et al.*, 2005]) : un k-PPV est directement défini à partir de tous les exemples de l'échantillon d'apprentissage sans information supplémentaire. Alors que d'autres classifieurs, comme les *decision list machines* [Marchand et Sokolova, 2005], requièrent une séquence de compression ainsi qu'un message. Nous donnons dans la suite une borne en généralisation valide pour tout schéma de compression.

Bornes en généralisation dans le cas de votants dépendants des données

Soit $S_{\mathbf{i}}$ une séquence de compression composée de $|\mathbf{i}|$ exemples issus de l'ensemble d'apprentissage S. Dans le contexte d'un schéma de compression PAC-Bayésien, les erreurs $\mathbf{R}_P(\cdot)$ et $\mathbf{R}_S(\cdot)$ peuvent être biaisées par ces éléments : il est donc préférable de calculer l'erreur empirique $\mathbf{R}_S(\cdot)$ à partir de $S \backslash S_{\mathbf{i}}$ [Laviolette et Marchand, 2007]. Cependant, pour dériver une borne sur l'erreur dans une telle situation, [Germain *et al.*, 2011] ont proposé une stratégie différente pour prendre en compte le biais. En suivant cette stratégie et étant donné un échantillon d'apprentissage S, nous considérons $\mathcal{H}^S$ l'ensemble de tous les classifieurs possibles $h^{\omega}_{S_{\mathbf{i}}} = \mathcal{R}(S_{\mathbf{i}}, \omega)$ tel que $\omega \in \Omega_{S_{\mathbf{i}}}$. Nous notons $\rho_{\mathbf{I}_m}(\mathbf{i})$ la probabilité qu'une séquence de compression $S_{\mathbf{i}}$ soit choisie par ρ, et $\rho_{S_{\mathbf{i}}}(\omega)$ la probabilité de choisir un message ω sachant $S_{\mathbf{i}}$. Alors :

$$\rho_{\mathbf{I}_m}(\mathbf{i}) = \int_{\omega \in \Omega_{S_{\mathbf{i}}}} \rho(h^{\omega}_{S_{\mathbf{i}}}) d\omega, \quad \text{et} \quad \rho_{S_{\mathbf{i}}}(\omega) = \rho(h^{\omega}_{S_{\mathbf{i}}} | S_{\mathbf{i}}).$$

Dans la théorie PAC-Bayésienne, les bornes sur l'erreur en généralisation dépendent de la distribution prior π sur l'ensemble $\mathcal{H}^S$. Ce prior est supposé connu avant l'observation de l'échantillon d'apprentissage S, impliquant que π doit être indépendant de S. Or, les votants de $\mathcal{H}^S$ dépendent de S et empêchent une telle connaissance *a priori*. Ce problème peut être contré, selon le principe de [Laviolette et Marchand, 2007, Germain *et al.*, 2011] en considérant une distribution prior définie par le couple :

$$\left(\pi_{\mathbf{I}_m}, (\pi_{S_{\mathbf{i}}})_{\mathbf{i} \in \mathbf{I}_m}\right),$$

où $\pi_{\mathbf{I}_m}$ est une distribution de probabilité sur l'ensemble d'indices $\mathbf{I}_m$ et $\pi_{S_{\mathbf{i}}}$ est une distribution de probabilité sur l'ensemble des messages $\Omega_{S_{\mathbf{i}}}$, pour toutes les séquences $S_{\mathbf{i}}$ possibles. Ainsi la distribution prior π indépendante des données de S correspond à la distribution sur $\mathcal{H}^S$ associée au prior $(\pi_{\mathbf{I}_m}, (\pi_{S_{\mathbf{i}}})_{\mathbf{i} \in \mathbf{I}_m})$ et est définie par :

$$\forall \mathbf{i} \in \mathbf{I}_m,\ \forall \omega \in \Omega_{S_{\mathbf{i}}},\ \pi(h^{\omega}_{S_{\mathbf{i}}}) = \pi_{\mathbf{I}_m} \pi_{S_{\mathbf{i}}}(\omega).$$

Dans cette situation, la ρ-marge du vote de majorité ρ-pondéré est définie par :

Définition 4.2 *Dans le cas d'un schéma de compression, la ρ-marge de $B_\rho(\cdot)$ mesurée sur un exemple $(\mathbf{x}, y)$:*

$$\mathcal{M}^\rho(\mathbf{x}, y) = y \mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{i}} \sim \rho} h^\omega_{S_\mathbf{i}}(\mathbf{x}).$$

Soit P un domaine sur $X \times Y$. Soit un échantillon $S \sim (P)^m$. Les premiers moments réel $\mathcal{M}^\rho_P$ et empirique $\mathcal{M}^\rho_S$ et les seconds moments réel $\mathcal{M}^{\rho^2}_P$ et emirique $\mathcal{M}^{\rho^2}_S$ de la ρ-marge sont définis comme précédemment par :

$$\mathcal{M}^\rho_P = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathcal{M}^\rho(\mathbf{x}, y), \quad et \quad \mathcal{M}^\rho_S = \frac{1}{m}\sum_{i=1}^m \mathcal{M}^\rho(\mathbf{x}_i, y_i),$$
$$\mathcal{M}^{\rho^2}_P = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y)\right)^2, \quad et \quad \mathcal{M}^{\rho^2}_S = \frac{1}{m}\sum_{i=1}^m \left(\mathcal{M}^\rho(\mathbf{x}_i, y_i)\right)^2.$$

Rappelons que nous considérons une famille auto-complémentée de votants $\mathcal{H}^S$ et uniquement des distributions de probabilité π-alignées sur $\mathcal{H}^S$. Pour tout votant h^ω_S dans l'ensemble $\mathcal{H}^S$, son opposé est noté $-h^\omega_S$. Ainsi, étant donné un échantillon S, l'ensemble des messages associé est défini par $\Omega_S \times \{+,-\}$ et : $\forall \sigma \in \Omega_S$, $h^{(\sigma,+)}_S = -h^{(\sigma,-)}_S$. Le résultat principal de cette section est donné dans le théorème suivant.

Théorème 4.1 *Soit P un domaine sur $X \times Y$, soit $m \geq 8$, soit S un échantillon de m éléments i.i.d. selon P. Alors pour tout $\delta \in (0, 1]$, avec une probabilité d'au moins $1 - \delta$ sur le choix de $S \sim (P)^m$ (de taille m), pour tout ensemble $\mathcal{H}^S$ auto-complémenté de votants bornés par B et de taille de séquence de compression au plus $|\mathbf{i}_{\max}| < \frac{m}{2}$ et pour toute distribution π-alignée ρ sur l'ensemble $\mathcal{H}^S$, on a :*

$$\left|\mathcal{M}^\rho_P - \mathcal{M}^\rho_S\right| \leq \frac{2B}{\sqrt{2(m - |\mathbf{i}_{\max}|)}} \sqrt{\frac{|\mathbf{i}_{\max}|}{B\delta} + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}, \tag{4.2}$$

et :

$$\left|\mathcal{M}^{\rho^2}_P - \mathcal{M}^{\rho^2}_S\right| \leq \frac{2B^2}{\sqrt{2(m - 2|\mathbf{i}_{\max}|)}} \sqrt{\frac{2|\mathbf{i}_{\max}|}{B^2\delta} + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}. \tag{4.3}$$

Démonstration. En annexe C.2 (inspirée de la preuve du théorème 3.4, chapitre 3). □

Cette borne est à la fois valide pour MinCq et pour sa version π-alignée.
Notons que dans le cas de classifieurs indépendants des données d'apprentissage, c'est-à-dire lorsque $|\mathbf{i}_{\max}| = 0$, on retrouve le théorème 3.4. Comme attendu, plus les séquences de compression sont grandes, *i.e.* plus $|\mathbf{i}_{\max}|$ est élevé, moins la borne est précise. Ainsi, pour préserver la consistance du processus d'apprentissage, cette taille ne doit pas être trop importante.
Nous allons maintenant étudier deux instanciations concrètes de l'approche MinCq/P-MinCq. La première, en section 4.2, propose une approche originale pour combiner des k-PPV avec différentes valeurs de k. Nous proposons un *a priori* naturel à chaque votant qui nous permettra de souligner l'intérêt du π-alignement. La seconde, en section 4.3, se focalise sur un vote de majorité basé sur des votants appris à partir de différentes descriptions des données. Plus précisément, dans le contexte d'une problématique de

fusion de classifieurs en multimédia, nous présentons une régularisation spécifique à la tâche d'indexation de données avec une perspective de recherche d'information. Cette spécialisation met en évidence l'intérêt de l'approche MinCq/P-MinCq dans un tel contexte, mais pour lequel il est complexe de définir un π-alignement pertinent.

4.2 Application à des classifieurs de type k-PPV

4.2.1 Motivation

Nous rappelons qu'il existe deux types de stratégies pour améliorer les k-PPV (voir section 1.4.1 pour plus de détails). La première stratégie adapte localement les voisinages [Hastie et Tibshirani, 1996, Nock *et al.*, 2003]. La seconde se focalise sur l'apprentissage de métriques pour optimiser les distances entre les points de même classe. Quelle que soit la stratégie, le nombre de voisins k doit convenablement être sélectionné et la règle de décision, se basant sur les voisinages locaux, ne prévient pas du sur-apprentissage. Au contraire, l'approche MinCq/P-MinCq peut être utilisé pour optimiser un vote de majorité pondéré de classifieurs k-PPV, pour lequel l'ensemble des votants $\mathcal{H}$ serait constitué de classifieurs k-PPV (pour $k = \{1, 2, \ldots\}$). Comme nous le verrons dans une série d'expériences préliminaires, MinCq s'avère moins performant qu'un simple k-PPV. La raison de ce comportement vient du fait que la contrainte de quasi-uniformité nécessaire à MinCq suppose que les votants ont *a priori* la même importance, ce qui n'est clairement pas le cas des k-PPV, notamment dans un contexte d'échantillon d'apprentissage de taille finie. Il est donc préférable de faire appel à une contrainte de π-alignement telle que plus k est faible (respectivement élevé), plus l'importance *a priori* du classifieur est élevée (respectivement faible), reflétant de la précision accrue des voisinages locaux.

4.2.2 Limitations de la contrainte de quasi-uniformité pour les k-PPV

Au premier abord, MinCq semble être une solution intéressante pour contrer les limitations liées à l'utilisation des k-PPV.

- La théorie stipule que plus k est élevé, meilleure est la convergence vers le risque bayésien optimal. Cependant, ceci n'est vrai qu'asymptotiquement et, en pratique, le choix de k requiert une attention particulière[3]. Optimiser un vote de majorité de classifieurs k-PPV[4] ($k = \{1, 2, \ldots\}$) permettrait donc de s'affranchir du réglage de k.
- Les classifieurs k-PPV sont des classifieurs qui dépendent des données d'apprentissage et peuvent être uniquement reconstruits à partir d'une séquence de compression encodant les PPV. Ainsi, la minimisation empirique de la C-borne du

3. Voir section 1.4.1 pour plus de détails concernant le choix de k.

4. D'autres ensembles de votants pourraient être considérés. Par exemple, $n^{\text{ème}}$ voisin pourrait correspondre au $n^{\text{ième}}$ votant.

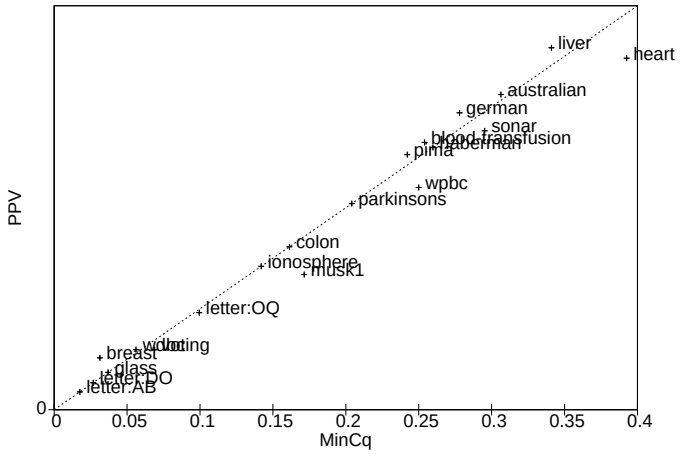


FIGURE 4.1 – *Comparaison de MinCq versus PPV. Un point correspond au taux d'erreur test des deux algorithmes sur un jeu de données. Un jeu de donnée au-dessus de la bissectrice est en faveur de MinCq.*

théorème 3.3 justifiée par les bornes PAC-Bayésienne de notre théorème 4.1, offre des garanties en généralisation qui ne peuvent être obtenues avec un algorithme de k-PPV standard dans le cas d'échantillons finis.

En suivant le protocole de la section 4.2.4, nous avons réalisé une étude expérimentale préliminaire afin de comparer un classifieur k-PPV classique (pour lequel, k a été choisi par validation croisée) avec MinCq cherchant un vote de majorité sur l'ensemble des k-PPV. Sur un ensemble de 20 jeux de données issus du *UCI Machine Learning Repository* [5]. MinCq atteint une erreur moyenne de 18.18% contre 17.88% pour le k-PPV (voir la table 4.2 pour plus de détails). Un test de Student ne distingue pas les deux approches. Ceci se confirme également par un test de signe avec un résultat *win/loss/tie* égal à 7/6/7 menant à une p-valeur d'environ 0.5. La figure 4.1 illustre ce comportement : aucune des deux approches n'est significativement meilleure que l'autre. Cette série d'expériences montre clairement que MinCq n'est pas meilleur qu'un k-PPV avec un k correctement choisi.

Ce résultat *a priori* surprenant peut en fait être expliqué par l'hypothèse non informative de quasi-uniformité faite sur ρ : cette hypothèse ne s'avère pas pertinente lorsque les votants n'ont pas la même importance. En effet, pour des raisons évidentes, un voisinage proche offre naturellement une meilleure information discriminante qu'une règle de décision des k-PPV avec un k très grand (dans le cas d'échantillon finis). Nous proposons donc par la suite d'instancier P-MinCq pour répondre à cette contrainte.

4.2.3 Instanciation de P-MinCq pour les k-PPV

Quelques mots sur la borne en généralisation

Rappelons que les classifieurs de type k-PPV sont dépendants des données d'apprentissage. Les garanties en généralisation sont donc données [6] par notre théorème 4.1.

5. http://archive.ics.uci.edu/ml/
6. Dans ce contexte les votants sont majorés par $B = 1$.

Cependant, quelle que soit la valeur de k, pour un échantillon d'apprentissage de taille m, la taille de la séquence de compression vaut m rendant invalide les bornes puisque les bornes sont valides lorsque $|\mathbf{i}_{\max}| \leq \frac{m}{2}$. Pour éviter cette forme indéterminée, $|\mathbf{i}_{\max}|$ peut considérablement être réduit par des techniques de sélection de prototypes ou de réduction de bases [Duda *et al.*, 2001]. Ces techniques permettent de supprimer de la séquence de compression les exemples ne modifiant pas la décision sur les données de test. Dans ce cas, chaque classifieur k-PPV utilise sa propre séquence de compression : un sous-ensemble de l'échantillon d'apprentissage de faible taille.

Une contrainte π statistiquement fondée

P-MinCq offre un contexte original pour les k-PPV, en combinant différents k-PPV. Au lieu de régler k, nous définissons une contrainte *a priori* de $\boldsymbol{\pi}$-alignement sur les votants. Comme mentionné par [Devroye *et al.*, 1996] ce sont les voisinages les plus proches qui apportent le plus d'information dans une combinaison de k-PPV. En suivant cette recommandation, nous proposons la contrainte $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ suivante, qui à une normalisation près correspond à une densité :

$$\forall k \geq 1, \quad \pi_k = \frac{1}{k}. \tag{4.4}$$

$\boldsymbol{\pi}$ concentre ses poids sur les votants définis à partir d'une faible fraction d'exemples[7], mais prend aussi en compte, dans une moindre mesure, l'information portée par l'ensemble des voisinages. Dans ce qui suit, nous justifions ce choix en établissant une relation entre l'équation (4.4) et sa médiane M (le rang k accumulant la moitié de la densité).

Alors que dans le cas d'une distribution continue le calcul de M est aisé, le cas discret (qui nous intéresse) requiert une approximation.

Soit :

$$H_M = \sum_{x=1}^{M} \frac{1}{x} \quad \text{et} \quad H_m = \sum_{x=1}^{m} \frac{1}{x}\,,$$

la somme des termes de séries harmoniques. Si H_M et H_m n'admettent aucune solution analytique, des sommes partielles de séries nous permettent, cependant, de contourner le problème :

$$\begin{aligned} \forall n,\ H_n &= \sum_{x=1}^{n} \frac{1}{x} \\ &= \ln(n) + \gamma + \epsilon_n, \end{aligned}$$

où $\gamma \simeq 0.5772156$ est la constante d'Euler-Mascheroni[8] et $\epsilon_n \simeq \frac{1}{2n}$.

7. Après application d'une technique de sélection de prototypes.

8. La constante d'Euler-Mascheroni est définie comme étant la limite de la différence entre la série harmonique et le logarithme naturel.

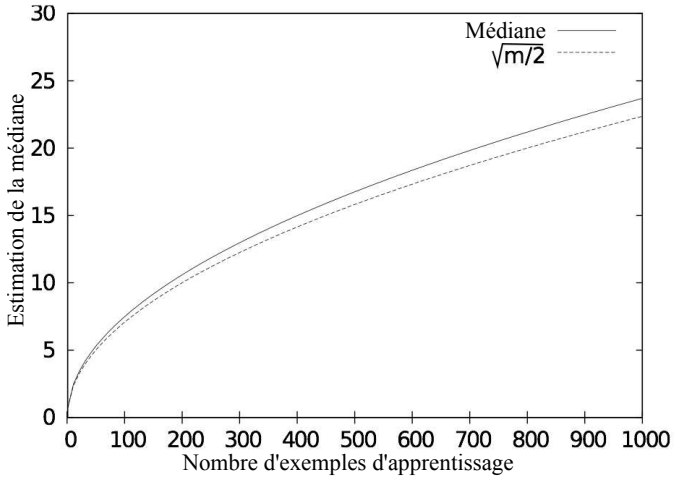


FIGURE 4.2 – *Comparaison de la médiane des séries harmoniques* $\sum_{x=1}^{m} \frac{1}{x}$ *et* $\sqrt{m/2}$.

Par conséquent :

$$
\begin{aligned}
H_M = \frac{1}{2} H_m &\iff \sum_{x=1}^{M} \frac{1}{x} = \frac{1}{2} \sum_{x=1}^{m} \frac{1}{x} \\
&\iff \ln(M) + \gamma + \frac{1}{2M} = \frac{1}{2}(\ln(m) + \gamma) + \frac{1}{4m} \\
&\iff \ln(M) = \ln(\sqrt{m}) - \frac{1}{2}\gamma + \frac{1}{4m} - \frac{1}{2M} \\
&\iff \ln(M) \leq \ln(\sqrt{m}) - \frac{1}{2}\gamma - \frac{1}{4m} \\
&\iff M \leq \sqrt{m \times \exp(-\gamma) \times \exp\left(-\frac{1}{4m}\right)} \simeq \sqrt{\frac{m}{2}}. \qquad (4.5)
\end{aligned}
$$

Dans le cas discret, l'équation (4.5) indique donc que l'approximation de la médiane de π est très proche de $\sqrt{m/2}$, la valeur de k suggérée pour la règle des k-PPV en classification binaire. La figure 4.2 l'illustre graphiquement. Nous avons donc établi une relation étroite entre la sélection classique de k et notre contrainte π. La section suivante présente une étude expérimentale permettant de valider ce choix de π.

4.2.4 Expérimentations

Dans cette section, nous réalisons une étude de P-MinCq pour combiner l'ensemble de k-PPVpour k allant de 1 à $\sqrt{m}$ comme décrit dans la section 4.2.3. Notons que pour chaque k-PPV nous avons appliqué au préalable une technique de sélection de prototypes. Nous nous comparons à quatre méthodes.

- L'algorithme standard des k-PPV, présenté dans la section 1.4.1 du chapitre 1, joue le rôle de référence (PPV).
- L'algorithme de k-PPV symétriques (SNN : *Symmetric Nearest Neighbor* [Nock *et al.*, 2003]), une variante de PPV où la classe d'un exemple $\mathbf{x}$ est déterminée par la classe majoritaire parmi les échantillons d'apprentissage appartenant au k-voisinage de $\mathbf{x}$ (comme pour k-PPV) ainsi que ceux incluant $\mathbf{x}$ dans leur propre k-voisinage.

Nom	Nombre d'attributs	Taille de S	Taille en test
australian	14	345	345
blood	4	374	374
breast	9	349	350
colon	2000	31	31
german	24	500	500
glass	9	107	107
haberman	3	153	153
heart	13	135	135
ionosphere	34	175	176
letterAB	16	297	1192
letterDO	16	297	1193
letterOQ	16	291	1166
liver	6	172	173
musk1	166	238	238
parkinsons	22	97	98
pima	8	384	384
sonar	60	104	104
voting	16	217	218
wdbc	30	284	285
wpbc	33	99	99

TABLE 4.1 – *Propriétés des* 20 *jeux de données considérés.*

- L'algorithme de *Large Margin Nearest Neighbor* (LMNN) [Weinberger et Saul, 2009] qui apprend essentiellement une distance de Mahalanobis en optimisant l'erreur du k-PPV sur l'ensemble d'apprentissage avec une marge de sécurité. Puis, un k-PPV est appliqué avec la distance apprise.
- L'algorithme MinCq [Laviolette *et al.*, 2011a] qui considère une distribution quasi-uniforme.

Nous évaluons tout d'abord ces approches sur 20 jeux de données de référence. Puis, nous nous attaquons à une tâche de reconnaissance d'objets dans des images.

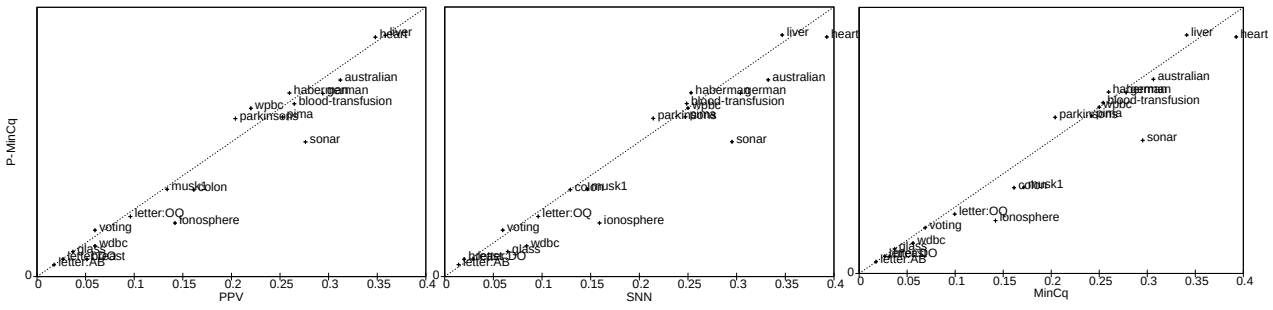
Jeux de données de référence

Protocole Les 20 jeux de données variés sont issus du *UCI Machine Learning Repository* [9] (voir la table 4.1 pour leurs caractéristiques). La distance euclidienne est utilisée pour calculer les voisinages. Les données sont découpées en 50% d'apprentissage et 50% de test, sauf *letterAB, letterDO* et *letterOQ* que nous divisons en 20%/80%. Les paramètres suivants sont sélectionnés par validation croisée sur 10 sous-ensembles de l'échantillon d'apprentissage : la marge μ de MinCq et P-MinCq (parmi 14 valeurs dans $[10^{-4}, 0.5]$) et le k des k-PPV et LMNN (parmi $\{1, \ldots, 10\}$). Le paramètre de compromis de LMNN est fixé à 0.5 comme dans [Weinberger et Saul, 2009].

9. `http://archive.ics.uci.edu/ml/`

Jeu de données	PPV	SNN	LMNN	MinCq	P-MinCq
australian	0.3121	0.3324	0.2746	0.3064	0.2919
blood	0.2647	0.2487	0.2674	0.2540	0.2567
breast	0.0514	0.0200	0.0400	0.0314	0.0257
colon	0.1613	0.1290	0.2258	0.1613	0.1290
german	0.2940	0.3040	0.2760	0.2780	0.2720
glass	0.0370	0.0648	0.0648	0.0370	0.0370
haberman	0.2597	0.2532	0.2922	0.2597	0.2727
heart	0.3481	0.3926	0.2148	0.3926	0.3556
ionosphere	0.1420	0.1591	0.1193	0.1420	0.0795
letter :AB	0.0176	0.0143	0.0151	0.0176	0.0176
letter :DO	0.0268	0.0293	0.0126	0.0268	0.0260
letter :OQ	0.0961	0.0961	0.0334	0.0995	0.0892
liver	0.3584	0.3468	0.3584	0.3410	0.3584
musk1	0.1339	0.1464	0.2092	0.1715	0.1297
parkinsons	0.2041	0.2143	0.1531	0.2041	0.2347
pima	0.2526	0.2474	0.2604	0.2422	0.2370
sonar	0.2762	0.2952	0.0762	0.2952	0.2000
voting	0.0596	0.0596	0.0413	0.0688	0.0688
wdbc	0.0596	0.0842	0.0491	0.0561	0.0456
wpbc	0.2200	0.2500	0.2300	0.2500	0.2500
Erreur moyenne	0.1788	0.1844	0.1607	0.1818	0.1689
Rang moyen	2.9	3.1	2.65	2.9	2.25

TABLE 4.2 – *Taux d'erreur de PPV, SNN, LMNN, MinCq et P-MinCq sur les* 20 *jeux de données.*



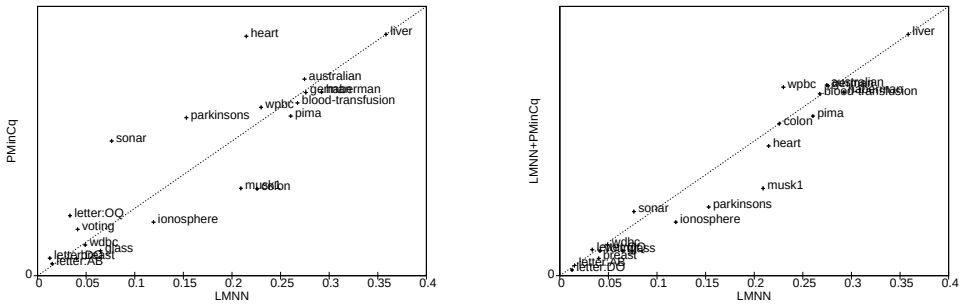
(a) *P-MinCq versus PPV* **(b)** *P-MinCq versus SNN* **(c)** *P-MinCq versus MinCq*

FIGURE 4.3 – *Comparaison de P-MinCq, PPV, SNN et MinCq. Un point correspond au taux d'erreur des deux algorithmes comparés. Un jeu de données en dessous de la bissectrice est en faveur de P-MinCq.*

Résultats Les résultats obtenus sur les ensembles de test sont reportés dans la table 4.2. P-MinCq se montre plus performant qu'un PPV classique. En moyenne, P-MinCq atteint un taux d'erreur de 16.89% contre 17.88% pour PPV. Avec un test de Student apparié, cette différence est statistiquement significative avec une p-valeur de 0.06. Ce résultat est conforté par un test de signe renvoyant un résultat *win/loss/tie* égal à 12/5/3 avec une p-valeur de 0.07 (figure 4.3(a)). De plus, la figure 4.3(b) montre que P-MinCq est meilleur que SNN (qui est plus performant sur quelques jeux de données) : une p-valeur de 0.01 en faveur de P-MinCq avec un test de Student et de 0.24 avec un test de signe. En outre, P-MinCq améliore MinCq : un test de Student mène à une p-valeur de 0.02, un test de signe à une p-valeur $\simeq 0.03$ avec un *win/loss/tie* de 12/4/4 (voir la figure 4.3(c)). Cette étude montre, d'une part, l'intérêt de la généralisation de MinCq aux distributions π-alignées et, d'autre part, la pertinence de la contrainte $\pi_k = \frac{1}{k}$ (normalisée) pour les PPV. Même s'il n'est pas un algorithme d'apprentissage

Jeu de données	LMNN	LMNN+P-MinCq
australian	0.2746	0.2832
blood	0.2674	0.2701
breast	0.0400	0.0257
colon	0.2258	0.2258
german	0.2760	0.2820
glass	0.0648	0.0370
haberman	0.2922	0.2727
heart	0.2148	0.1926
ionosphere	0.1193	0.0795
letter :AB	0.0151	0.0151
letter :DO	0.0126	0.0084
letter :OQ	0.0334	0.0386
liver	0.3584	0.3584
musk1	0.2092	0.1297
parkinsons	0.1531	0.1020
pima	0.2604	0.2370
sonar	**0.0762**	0.0952
voting	0.0413	0.0367
wdbc	0.0491	0.0456
wpbc	**0.2300**	0.2800
Erreur moyenne	0.1607	**0.1508**

TABLE 4.3 – *Taux d'erreur de LMNN et LMNN+P-MinCq sur les* 20 *jeux de données.*



(a) *P-MinCq versus LMNN* **(b)** *P-MinCq+LMNN versus LMNN*

FIGURE 4.4 – *Comparaison de P-MinCq et LMNN et de P-MinCq+LMNN et LMNN.*

de métrique, P-MinCq s'avère compétitif avec LMNN (0.1689 contre 0.1607 d'erreur moyenne avec une p-valeur d'environ 0.10 pour un test de Student). Un test de signe amène à une p-valeur de 0.5, indiquant qu'aucune des méthodes n'est meilleure que l'autre. La figure 4.4(a) montre que P-MinCq et LMNN sont plutôt complémentaires. LMNN apprend une métrique qui ajuste les voisinages (parfois avec un grand succès, par exemple *heart*, *parkinsons*, *sonar*), mais est parfois moins performant que PPV car une dimensionnalité élevée implique un fort risque de sur-apprentissage (par exemple : *colon*, *musk1*) comme mentionné dans [Bellet *et al.*, 2012]. P-MinCq, quant à lui, combine différentes règles de PPV et n'intervient donc pas sur les voisinages : cette combinaison de votants apparaît plus stable (voir la table 4.2 avec le meilleur rang moyen) et plus robuste au sur-apprentissage. Afin de mesurer cette complémentarité, nous réalisons une série d'expériences complémentaire dont le but est de combiner LMNN et P-MinCq (lorsque cela paraît pertinent). Concrètement, si LMNN est plus pertinent

FIGURE 4.5 – *Quelques exemples de bikes (colonne de gauche), persons (milieu) et d'arrière plans (droite) issus de Graz-01. Uniquement des parties des objets peuvent être visible, la classe d'arrière plan est difficile à identifier, par exemple : il est difficile de distinguer un vélo d'une moto.*

que P-MinCq sur l'ensemble de validation, la distance apprise par LMNN est privilégiée par P-MinCq (sinon la distance euclidienne est conservée). Nous reportons les résultats dans la table 4.3. La combinaison LMNN+P-MinCq bat clairement toutes les autres méthodes, y compris LMNN seul (un test de Student avec une p-valeur de 0.05 et un test de signe à 0.17), comme illustré par la figure 4.4(b). Notons que sur les jeux de données sur lesquels LMNN était le plus performant (par exemple *heart*, *parkinson*, *voting*), LMNN+P-MinCq est capable d'améliorer encore plus ces résultats.

Reconnaissance d'objets

Protocole Nous réalisons une expérience sur Graz-01 [Opelt *et al.*, 2004], un jeu de données où deux classes d'objets sont à identifier (*bike*, *person*), ainsi qu'une classe d'arrière plan. Graz-01 est connu pour sa grande variation intra-classes et ses arrières plans très parasités (voir la figure 4.5). Les tâches de classification sont : *bike*/non-*bike* et *person*/non-*person*. Nous suivons le protocole de [Opelt *et al.*, 2004] : pour chaque objet, 100 images positives et 100 négatives sont choisies aléatoirement (50 issues de l'autre objet et 50 de l'arrière plan). Les images sont décrites par un histogramme de fréquences de 200 mots visuels construits à partir des points d'intérêts SIFT [Lowe, 1999]. Les voisinages sont calculés via deux distances d'histogrammes [10] la distance χ^2 et la distance d'intersection.

Résultats Nous reportons dans la table 4.4 les résultats moyennés sur 10 tirages aléatoires. P-MinCq est encore une fois le plus stable et le meilleur en moyenne. Il est plus performant que PPV et MinCq (une p-valeur inférieure à 0.01 avec un test de Student) et que SSN dans une moindre mesure (une p-valeur de 0.13). Notons que SNN améliore significativement PPV sur ces données : la variation intra-classe semble rendre payante l'extension du voisinage. Cependant, alors que l'heuristique de

10. Si H_1 et H_2 sont deux histogrammes de même taille n, la distance χ^2 est définie par : $\sum_{i=1}^{n} \frac{(H_1(i)-H_2(i))^2}{H_1(i)-H_2(i)}$, et la distance d'intersection par : $\sum_{i=1}^{n} \min\{H_1(i), H_2(i)\}$.

Distance	Tâche	PPV	SNN	MinCq	P-MinCq
χ^2	bike	0.2310	0.2090	0.2160	0.2095
χ^2	person	0.2385	0.2305	0.2730	0.2250
Intersection	bike	0.2260	0.2185	0.2130	0.2055
Intersection	person	0.2350	0.2370	0.3180	0.2255
Erreur moyenne		0.2326	0.2238	0.2550	0.2164

TABLE 4.4 – *Taux d'erreur moyennés sur* 10 *tirages de PPV, SNN, MinCq et P-MinCq sur Graz-01.*

symétrie de SNN n'est pas toujours pertinente (comme pour les jeux de données de référence), P-MinCq propose une alternative robuste et fondée théoriquement.

4.2.5 Conclusion

Nous avons proposé une nouvelle vision de la classification par k-PPV : l'objectif est d'apprendre un vote de majorité sur un ensemble de classifieurs k-PPV pour éviter, en partie, le réglage de k. Dans ce contexte, P-MinCq nous a permis de prendre en considération un *a priori* sur l'importance des différents k-PPV à combiner. Le π-alignement spécifique aux k-PPV modélise la plus grande influence des voisinages locaux. Les expériences menées dans cette section ont donc permis de mettre en évidence l'intérêt de la contrainte de π-alignement, lorsqu'il est possible d'en définir une. Cependant, comme nous allons le voir dans la section suivante, où nous spécialisons P-MinCq à la tâche de fusion de classifieurs, il est parfois difficile de définir un π-alignement pertinent.

4.3 Spécialisation à la fusion tardive de classifieurs

4.3.1 Motivation

P-MinCq est considéré comme un algorithme d'apprentissage d'un vote de majorité sur ensemble de votants qui peuvent être vus comme des fonctions de score [11] ou des régresseurs. Dans cet esprit, nous proposons d'étudier l'intérêt d'une approche de type MinCq pour combiner des classifieurs appris à partir de modalités différentes. Faire appel à plusieurs modalités permet de considérer des informations complémentaires issues de sources variées permettant d'améliorer la qualité d'un classifieur et correspondent généralement à un ensemble de descripteurs pertinents pouvant être regroupées en différentes vues ou modalités. Si n est le nombre de modalités, l'espace d'entrée est alors défini par : $X = X_1 \times \cdots \times X_n$. Dans cette section, nous nous concentrons sur une problématique d'indexation de données multimédia [12] où l'on parle alors

11. La fonction de score associée à un modèle est la fonction renvoyant la valeur réelle avant d'en prendre le signe.

12. Pour un état de l'art voir [Atrey *et al.*, 2010].

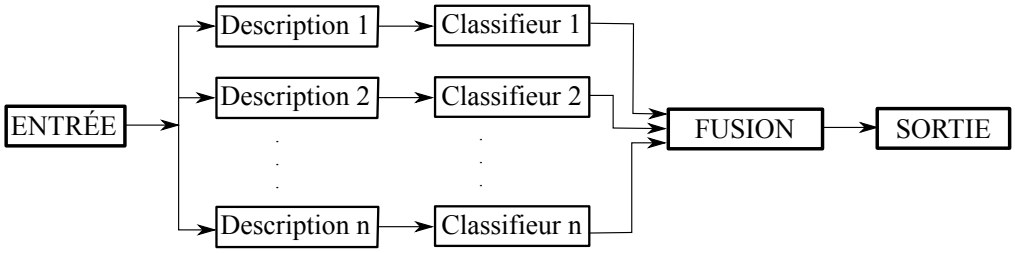


FIGURE 4.6 – *Principe de la fusion tardive de classifieurs appris à partir de n descriptions différentes.*

de fusion de classifieurs. Dans ce contexte particulier, deux approches sont principalement considérées [Snoek *et al.*, 2005] : la fusion précoce et la fusion tardive. Concernant la fusion précoce, les descriptions sont fusionnées en un seul vecteur avant la phase d'apprentissage. Ceci peut être vu comme une classification unimodale. Cependant, ce genre d'approche est parfois difficile à appréhender car les données/attributs sont souvent nombreux et très hétérogènes. La fusion tardive (voir la figure 4.6) vise, quant à elle, à combiner les scores de prédictions lors de la prise de décision. Ce principe est généralement appelé classification multimodale ou fusion de classifieurs. Cette approche ne montre pas nécessairement de meilleurs résultats qu'une classification unimodale, en particulier, lorsqu'une modalité permet d'apprendre un classifieur significativement plus performant que ceux appris à partir des autres modalités. Néanmoins, la fusion tardive tend à être plus performante dans des situations d'apprentissage de concepts sémantiques dans des vidéos multimodales [Snoek *et al.*, 2005]. De nombreuses méthodes sont basées sur une règle de décision fixe telle que le score maximal ou minimal, le produit ou la somme des scores,... [Kittler *et al.*, 1998]. D'autres approches, appelées *stacking*, requièrent une phase d'apprentissage supplémentaire [Wolpert, 1992, Atrey *et al.*, 2010, Kuncheva, 2004, Dietterich, 2000].

Nous considérons ici le problème de la fusion tardive des fonctions de score de classifieurs appris sur différentes modalités. On note $h_j : X_j \mapsto \mathbb{R}$ la fonction de score associée à la $j^{\text{ième}}$ modalité X_j. Une méthode classique consiste à apprendre une combinaison linéaire des différents scores, pouvant être vue comme un vote de majorité pondéré :

$$B_\rho^{reg}(\mathbf{x}) = \sum_{j=1}^{n} \rho_j h_j(\mathbf{x}),$$

où ρ_j représente le poids associé à h_j (pour rester cohérent avec la théorie PAC-Bayésienne, on suppose $0 \leq \rho_j \leq 1$ et $\sum_{j=1}^{n} \rho = 1$) et $B_\rho^{reg}(\cdot)$ la fonction score associée au vote de majorité classique B_ρ (c'est-à-dire sans la fonction $\text{sign}(\cdot)$). Lorsque l'on apprend une telle combinaison, il est important de considérer un ensemble de votants suffisamment diverses [Dietterich, 2000]. Par exemple, l'algorithme AdaBoost [Freund et Schapire, 1996], très souvent utilisé comme méthode de fusion multimodale, pondère les votants selon différentes distributions sur les données d'apprentissage en introduisant une certaine diversité. Cependant, il requiert l'utilisation de classifieurs dits faibles (c'est-à-dire capables de discriminer deux classes légèrement mieux que le hasard) pour être performant : les auteurs de [Wickramaratna *et al.*, 2001] ont montré que la performance d'AdaBoost diminue si l'on combine des classi-

fieur forts comme, par exemple, des classifieurs-SVM. Une autre approche récente [Wang et Kankanhalli, 2010] se base sur la théorie du portefeuille en proposant une procédure de fusion minimisant les risques sur différentes modalités à l'aide d'une mesure de corrélation. Alors que cette approche est fondée théoriquement, elle nécessite la définition de fonctions appropriées et n'est pas complètement adaptée au problème puisqu'elle ne tient pas directement compte de la diversité des votants.
En gardant à l'esprit que MinCq/P-MinCq minimise la C-borne qui tient compte à la fois de l'erreur et du désaccord entre paires de votants, ou autrement dit, de la diversité (voir la section suivante), et offre de fortes garanties théoriques, il nous semble approprié de l'utiliser comme un algorithme de fusion de classifieurs multimodaux. MinCq/P-MinCq est ainsi capable d'inférer des combinaisons linéaires plus performantes que les combinaisons basiques, mais permet aussi l'utilisation d'une "couche supplémentaire" de noyaux en cherchant à apprendre un vote de majorité sur l'ensemble de ces noyaux. De plus, puisqu'en indexation sémantique de données multimédias, la mesure de performance est très souvent reliée au classement des exemples positifs, nous proposons d'étendre P-MinCq en optimisant la mesure *Mean Average Precision* (MAP). Cette extension fait appel à une perte additionnelle préservant l'ordre pour vérifier des contraintes de rang par paires.

4.3.2 P-MinCq vu comme un algorithme de fusion de classifieurs

Justification de l'utilisation de l'approche P-MinCq

Nous justifions, tout d'abord, de l'utilisation de l'approche P-MinCq dans ce contexte. Pour que les bornes en généralisation du théorème 4.1 soient valides, nous allons suivre le protocole suivant. Étant donné un échantillon d'apprentissage de taille $2m$, nous le découpons aléatoirement en deux sous-ensembles S' et S de même taille m. Nous supposons que nos données sont décrites selon n modalités/descriptions différentes. Pour chaque modalité j, nous apprenons un régresseur h_j depuis S'. On considère alors $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$ l'ensemble auto-complémenté de ces fonctions de score et les opposées : $h_{j+n} = -h_j$. Si l'on suppose une connaissance *a priori* sur la pertinence des modalités, nous pouvons la modéliser par une contrainte $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$. Dans le cas contraire, nous ferons appel à la contrainte de quasi-uniformité. À ce stade, la fusion est donc réalisée par P-MinCq : le vote de majorité ρ-pondéré est appris depuis l'ensemble $S = \{(\mathbf{x}_i, y_i)\}_{j=1}^{m}$.

Un autre point important concerne la notion de diversité qui est un élément clé pour combiner avec succès des fonctions [Dietterich, 2000, Kuncheva, 2004, Atrey *et al.*, 2010]. Du point de vue de l'indexation multimédia, la fusion de votants diverses est nécessaire pour espérer de bonnes performances. Il est intéressant de souligner que P-MinCq favorise les votes de majorités pour lesquels les votants sont les plus décorrélés. Nous expliquons un peu plus en détail ce comportement.
Dans la littérature, il n'existe pas de définition générale de la diversité. Cependant,

plusieurs métriques de diversité classiques se basent sur une comparaison de toutes les paires de classifieurs individuels, telles que les Q-statistiques, le coefficient de corrélation, la mesure de désaccord, ... [Kuncheva, 2004]. Ici, nous considérons la mesure de corrélation calculant le désaccord entre les prédictions de votants d'une même paire selon un domaine P :

$$\mathrm{div}_P(h_j, h_{j'}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} h_j(\mathbf{x})h_{j'}(\mathbf{x}).$$

Nous pouvons réécrire le second moment de la ρ-marge (voir la définition 3.2) :

$$\begin{aligned}\mathcal{M}_D^{\rho^2} &= \sum_{j=1}^{2n}\sum_{j'=1}^{2n} \rho_j\rho_{j'}\,\mathrm{div}_P(h_j, h_{j'}) \\ &= \sum_{j=1}^{n}\sum_{j'=1}^{n} \left(2\rho_j - \pi_j\right)\left(2\rho_{j'} - \pi_{j'}\right)\mathrm{div}_P(h_j, h_{j'}). \qquad (4.6)\end{aligned}$$

La seconde ligne est obtenue grâce à l'auto-complémentation de $\mathcal{H}$ et au $\boldsymbol{\pi}$-alignement. L'approche P-MinCq minimise ce second moment $\mathcal{M}_D^{\rho^2}$ et optimise donc l'équation (4.6). Ceci implique une maximisation directe de la diversité entre votants d'une même paire : MinCq/P-MinCq favorise les votants les plus décorrélés et semble être une solution naturelle au problème de la fusion tardive de classifieurs appris séparément depuis des modalités différentes et variées.

Ajouter une contrainte d'ordonnancement

Dans de nombreuses applications, telles qu'en indexation de documents, l'objectif n'est pas d'étiqueter les données mais plutôt de les classer par ordre de pertinence (on parle souvent de *ranking*). Dans ce cas, les mesures d'évaluation traditionnellement utilisées sont reliées à la précision et au rappel. Un bon classifieur n'est pas nécessairement une bonne fonction d'ordonnancement. Nous proposons donc dans cette section une adaptation de P-MinCq pour aider à optimiser une des mesures les plus populaires en indexation et appelée MAP (*Mean Average Precision*). La mesure MAP calculée sur un échantillon S pour une fonction à valeur réelle h se définit comme suit. Soit $S^+ = \{(\mathbf{x}_i, y_i) : (\mathbf{x}_i, y_i) \in S \wedge y_i = 1\} = \{(\mathbf{x}_{i^+}, 1)\}_{i^+=1}^{m^+}$ l'ensemble des exemples positifs issus de S et $S^- = \{(\mathbf{x}_i, y_i) : (\mathbf{x}_i, y_i) \in S \wedge y_i = -1\} = \{(\mathbf{x}_{i^-}, -1)\}_{i^-=1}^{m^-}$ l'ensemble des exemples négatifs ($m^+ + m^- = m$). Pour évaluer le MAP, on trie les exemples par ordre décroissant de scores. Le MAP de h évalué sur S est défini par :

$$MAP_S(h) = \frac{1}{|m^+|}\sum_{i:y_i=+1} Prec@i,$$

où *Prec@i* est la proportion de positifs parmi les i premiers exemples. Ainsi, le score des exemples positifs doit être plus élevé que le score des exemples négatifs. Pour imposer un tel classement, nous proposons un apprentissage avec comparaisons par paires d'exemples positifs/négatifs [13] [Fürnkranz et Hüllermeier (eds), 2010, Zhang, 2004].

13. *Pairwise preference* en anglais.

En fait, les méthodes par paires offrent un compromis entre bonne classification et mesure de performance plus complexe telle que le MAP. Nous nous inspirons ici de la stratégie proposée par [Yue *et al.*, 2007] pour apprendre un classifieur- SVM en se basant sur une relaxation de la mesure de MAP. Dans la situation particulière de P-MinCq vu comme une méthode de fusion de classifieurs, nous avons développé une fonction de perte par couple préservant l'ordre. Concrètement, pour chaque couple $(\mathbf{x}_{i^+}, \mathbf{x}_{i^-}) \in S^+ \times S^-$, nous voulons vérifier :

$$B_\rho^{reg}(\mathbf{x}_{i^+}) > B_\rho^{reg}(\mathbf{x}_{i^-}) \iff B_\rho^{reg}(\mathbf{x}_{i^-}) - B_\rho^{reg}(\mathbf{x}_{i^+}) < 0. \tag{4.7}$$

Pour aider à la vérification de l'équation (4.7), nous proposons la relaxation suivante :

$$\frac{1}{m^+ m^-} \sum_{i^+=1}^{m^+} \sum_{i^-=1}^{m^-} \left[B_\rho^{reg}(\mathbf{x}_{i^-}) - B_\rho^{reg}(\mathbf{x}_{i^+}) \right]_+, \tag{4.8}$$

où nous rappelons que $\left[a\right]_+ = \max(0, a)$.
Dans le cadre général de P-MinCq, lorsque $\mathcal{H}$ est auto-complémenté et que la distribution ρ est π-alignée, l'équation (4.8), se réduit à :

$$\frac{1}{m^+ m^-} \sum_{i^+=1}^{m^+} \sum_{i^-=1}^{m^-} \left[\sum_{j=1}^{n} \left(2\rho_j - \pi_j\right) \left(h_j(\mathbf{x}_{i^-}) - h_j(\mathbf{x}_{i^+})\right) \right]_+ . \tag{4.9}$$

D'un point de vue algorithmique, pour minimiser la perte hinge de (4.9), nous faisons appel à $m^+ \times m^-$ variables additionnelles $\boldsymbol{\xi}_{S^+ \times S^-} = (\xi_{i^+ i^-})_{1 \leq i^+ \leq m^+, 1 \leq i^- \leq m^-}$ pondérées par un paramètre $\beta > 0$. Puisque $\boldsymbol{\xi}_{S^+ \times S^-}$ est un terme linéaire, par abus de notations et par souci de simplification, nous ajoutons l'équation (4.9) en tant que régularisateur à la fonction objectif du problème (4.1). Soit un échantillon d'apprentissage S constitué de m éléments tirés *i.i.d.* selon P, soit $\mathcal{H}$ un ensemble de votants auto-complémenté, soit $\mu > 0$, soit un vecteur *a priori* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$. Trouver la distribution posterior ρ sur $\mathcal{H}$ minimisant à la fois la C-borne et l'équation (4.9) est alors équivalent à trouver le vecteur de poids $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)^\top$, où $\rho_j = \rho(h_j)$ qui minimise le programme quadratique simple suivant :

$$\begin{cases} \min\limits_{\boldsymbol{\rho}} & (\boldsymbol{\rho} - \boldsymbol{\pi})^\top \mathbf{M}_S \boldsymbol{\rho} + \beta\, \mathbf{Id}^\top \boldsymbol{\xi}_{S^+ \times S^-}, \\ \text{s.c.} & \mathbf{m}_S{}^\top (2\boldsymbol{\rho} - \boldsymbol{\pi}) = \mu, \\ & \forall i^+ \in \{1, \ldots, m^+\},\ \forall i^- \in \{1, \ldots, m^-\},\ \xi_{i^+ i^-} \geq 0, \\ & \qquad \xi_{i^+ i^-} \geq \dfrac{1}{m^+ m^-} \displaystyle\sum_{j=1}^{n} \left(2\rho_j - \pi_j\right) \left(h_j(\mathbf{x}_{i^-}) - h_j(\mathbf{x}_{i^+})\right), \\ & \forall j \in \{1, \ldots, n\}, \quad 0 \leq \rho_j \leq \pi_j, \end{cases} \tag{4.10}$$

où $\mathbf{Id} = (1, \ldots, 1)^\top$ de taille $(m^+ \times m^-)$, $\mathbf{M}_S$ est la matrice $n \times n$ dont les éléments d'indices $(j, j') \in \{1, \ldots, n\}^2$ sont définis par :

$$\frac{1}{m} \sum_{i=1}^{m} h_j(\mathbf{x}_i) h_{j'}(\mathbf{x}_i)$$

et :

$$\mathbf{m}_S = \left(\frac{1}{m} \sum_{i=1}^{m} y_i h_1(\mathbf{x}_i), \ldots, \frac{1}{m} \sum_{i=1}^{m} y_i h_n(\mathbf{x}_i) \right)^\top .$$

En pratique, en incorporant une quantité quadratique de variables supplémentaires, le problème devient difficile à résoudre en grande dimension. Pour simplifier l'optimisation, nous relaxons les contraintes en les moyennant. En fait, puisque nous sommes intéressés à classer correctement et avec un grand score les exemples positifs, nous espérons que le score des exemples positifs soit plus élevé que la moyenne des prédictions pour les exemples négatifs. Le programme quadratique obtenu ne contient plus que m^+ variables additionnelles et est décrit dans le problème suivant (avec les même hypothèses et notations) :

$$\begin{cases} \min\limits_{\boldsymbol{\rho}} & (\boldsymbol{\rho}-\boldsymbol{\pi})^\top \mathbf{M}_S \boldsymbol{\rho} + +\beta\ \mathbf{Id}^\top \boldsymbol{\xi}_{S^+}, \\ \text{s.c.} & \mathbf{m}_S{}^\top (2\boldsymbol{\rho}-\boldsymbol{\pi}) = \mu, \\ & \forall i^+ \in \{1,\dots,m^+\},\ \xi_{i^+} \geq 0, \\ & \quad \xi_{i^+} \geq \dfrac{1}{m^+ m^-} \displaystyle\sum_{i^-=1}^{m^-} \sum_{j=1}^{n} \left(2\rho_j - \pi_j\right) \left(h_j(\mathbf{x}_{i^-}) - h_j(\mathbf{x}_{i^+})\right), \\ & \forall j \in \{1,\dots,n\}, \quad 0 \leq \rho_j \leq \pi_j, \end{cases} \tag{4.11}$$

où $\mathbf{Id} = (1,\dots,1)^\top$ de taille m^+, Les deux problèmes précédents contiennent un terme de régularisation supplémentaire : on va chercher le meilleur compromis entre une grande diversité et un bon classement.

4.3.3 Expérimentations sur PascalVOC'07

Protocole

Dans cette section, nous évaluons empiriquement l'intérêt de P-MinCq en tant qu'algorithme de fusion de classifieurs. Nous expérimentons les différentes approches sur la base de données PascalVOC'07 [Everingham *et al.*, 2007]. L'objectif est d'identifier 20 objets (concepts) différents. Le corpus est constitué de 10 000 images (50% en apprentissage et 50% en test). Pour la plupart des concepts, le ratio $+/-$ entre exemples positifs et négatifs est inférieur à 10% (les classes sont très déséquilibrées). Pour simplifier le problème, nous avons généré un ensemble d'apprentissage constitué de tous les exemples positifs et d'exemples négatifs indépendemment tirés tels que le ratio $+/-$ soit de $1/3$. Nous gardons l'ensemble de test original. En fait, notre objectif n'est pas de fournir les meilleurs résultats sur ce jeu de données, mais d'évaluer l'intérêt de P-MinCq lors de la phase de fusion tardive dans un contexte d'indexation multimédia.
Nous considérons neuf descriptions des données différentes : SIFT, LBP, Percepts, 2 HOG, 2 LCH et 2 CM :

- Les SIFT que nous considérons sont calculés à partir d'une grille dense, puis sont associés à un dictionnaire de 1 000 mots visuels générés avec l'algorithme des *k-means*.
- Les Percepts sont similaires aux SIFT, les mots visuels sont reliés aux classes sémantiques à un niveau local (voir la figure 6.8 dans le chapitre 6 pour plus de détails).

- Les LPB sont des motifs binaires locaux calculés sur une grille de 2×2 blocs (la dimension est de 1 024). L'opérateur LPB étiquette (avec un nombre décimal) les pixels d'une image en moyennant les (3×3)-voisinages de chaque pixel. Les LBP sont connus pour être invariants à tout changement monotone du niveau de gris.
- Les HOG sont des histogrammes de gradients calculés sur une grille de 4×3 blocs. Chaque case est définie comme la somme des gradients de magnitudes sur 50 orientations (la dimension est de 600). Les HOG sont connus pour être invariants à la translation et au changement d'échelle.
- Les LCH sont des histogrammes de couleurs locaux représentés par un histogramme 3D $3\times3\times3$ sur une grille de 8×6 ou 4×3 blocs.
- Les CM sont les deux premiers moments des couleurs dans l'espace RGB représentés sur une grille de 8×6 ou 4×3 blocs.

Nous apprenons un classifieur-SVM depuis chacune des descriptions à l'aide de la librairie LibSVM [Chang et Lin, 2001]. Le noyau considéré est le noyau gaussien classique défini dans l'équation (1.12). L'ensemble des votants à fusionner $\mathcal{H}$ est alors constitué des neuf fonctions de score réelles associées à ces classifieurs-SVM (P-MinCq considère implicitement leurs opposés). Nous ne disposons d'aucun *a priori* sur les différents votants et considérons donc uniquement des distributions quasi-uniforme sur $\mathcal{H}$. En effet, la définition d'une contrainte π-alignée dans ce cas est une problématique [14] en soit.

Dans une première série d'expérimentations, nous comparons les 3 algorithmes P-MinCq (problèmes (4.1), (4.10) et (4.11) notés respectivement P-MinCq, P-MinCq$_{PW}$ et P-MinCq$_{PWav}$) aux 4 méthodes basiques suivantes :

- La meilleure hypothèse de $\mathcal{H}$: $h_{best} = \underset{h_j\in\mathcal{H}}{\operatorname{argmax}}\ MAP_S(h_j)$.
- La fonction de prédiction avec la plus grande confiance :
$$best(\mathbf{x}) = \underset{h_j\in\mathcal{H}:j\in\{1,\dots,n\}}{\operatorname{argmax}} \big|h_j(\mathbf{x})\big|.$$
- La somme des scores (non-pondérés) : $\Sigma(\mathbf{x}) = \sum_{j=1}^{n} h_j(\mathbf{x})$.
- Le vote de majorité pondéré par la valeur des MAP individuels :
$$\Sigma_{MAP}(\mathbf{x}) = \sum_{j=1}^{n} \frac{MAP_S(h_j)}{\sum_{j'=1}^{n} MAP_S(h_{j'})} h_j(\mathbf{x}).$$

Dans une seconde série d'expériences, nous introduisons une information non linéaire en introduisant une couche de noyaux gaussiens. Autrement dit, nous augmentons artificiellement la taille de l'ensemble $\mathcal{H}$. Les exemples sont alors représentés par le vecteur

14. Nous avons essayé différentes stratégies qui se sont toutes révélées moins performante que la distribution quasi-uniforme.

concept	P-MinCq$_{PWav}$	P-MinCq$_{PW}$	P-MinCq	Σ	Σ_{MAP}	*best*	h_{best}
aeroplane	0.487	0.486	**0.526**	0.460	0.241	0.287	0.382
bicycle	0.195	0.204	**0.221**	0.077	0.086	0.051	0.121
bird	0.169	0.137	**0.204**	0.110	0.093	0.113	0.123
boat	0.159	0.154	0.159	0.206	0.132	0.079	**0.258**
bottle	0.112	**0.126**	0.118	0.023	0.025	0.017	0.066
bus	0.167	0.166	**0.168**	0.161	0.098	0.089	0.116
car	**0.521**	0.465	0.495	0.227	0.161	0.208	0.214
cat	**0.230**	0.219	0.220	0.074	0.075	0.065	0.116
chair	**0.257**	0.193	0.230	0.242	0.129	0.178	0.227
cow	0.102	0.101	**0.118**	0.078	0.068	0.06	0.101
diningtable	0.118	0.131	0.149	**0.153**	0.091	0.093	0.124
dog	**0.260**	0.259	0.253	0.004	0.064	0.028	0.126
horse	0.301	0.259	0.303	**0.364**	0.195	0.141	0.221
motorbike	0.141	0.113	0.162	**0.193**	0.115	0.076	0.130
person	**0.624**	0.617	0.604	0.001	0.053	0.037	0.246
pottedplant	**0.067**	0.061	0.061	0.057	0.04	0.046	0.073
sheep	0.067	**0.096**	0.069	0.128	0.062	0.064	0.083
sofa	0.204	**0.208**	0.201	0.137	0.087	0.108	0.147
train	0.331	0.332	**0.335**	0.314	0.164	0.197	0.248
tvmonitor	**0.281**	**0.281**	0.256	0.015	0.102	0.069	0.171
Moyenne	0.240	0.231	**0.243**	0.151	0.104	0.100	0.165

TABLE 4.5 – *Résultats en MAP obtenus sur l'échantillon test de PascalVOC'07.*

des scores de tous les votants, puis $\mathcal{H}$ devient l'ensemble des valeurs des noyaux sur l'échantillon S : chaque exemple $\mathbf{x}$ de l'échantillon S est alors vu comme un votant $K\big(\cdot, h(\mathbf{x})\big)$. Dans cette situation nous nous comparons à un classifieur-SVM avec un noyau gaussien. On indice avec ^{rbf} les méthodes faisant appel à cette couche.

Les différents hyperparamètres sont sélectionnés par validation croisée sur cinq sous-ensembles. Les performances sont évaluées en MAP sur l'ensemble de test et sont reportées dans la table 4.5 pour la première série et dans la table 4.6 pour la seconde.

Résultats

Notons tout d'abord que la performance de Σ_{MAP} est plus faible que Σ, suggérant que la performance des classifieurs individuels n'est pas liée à leur importance lors de l'étape de fusion. Cette observation s'est confirmée expérimentalement : en considérant cette performance MAP *a priori* comme une contrainte de π-alignement, nous n'avons obtenu qu'une perte de performance.

Pour la première expérimentation, la table 4.5 indique clairement que les approches P-MinCq sont en moyennes meilleures que les méthodes basiques. Elles renvoient la valeur du MAP la plus élevée pour 16 concepts sur 20. Ce résultat est confirmé statistiquement à l'aide d'un test de Student menant à une p-valeur proche de 0 au regard de Σ_{MAP}, *best* et h_{best}. En comparaison de Σ, les p-valeurs associées respectivement à P-MinCq$_{PWav}$, P-MinCq$_{PW}$ et P-MinCq sont 0.0139, 0.0232 et 0.0088. On peut remarquer que P-MinCq$_{PW}$ est moins performant que sa relaxation P-MinCq$_{PWav}$. Un test de Student mène à un p-valeur de 0.223, sans être significatif. Ainsi, lorsque notre objectif est de classer les exemples positifs avant les exemples négatifs, les

concept	P-MinCq$^{rbf}_{PWav}$	P-MinCq^{rbf}	SVM^{rbf}
aeroplane	**0.513**	**0.513**	0.497
bicycle	**0.273**	0.219	0.232
bird	**0.266**	0.264	0.196
boat	**0.267**	0.242	0.240
bottle	**0.103**	0.099	0.042
bus	**0.261**	**0.261**	0.212
car	**0.530**	**0.530**	0.399
cat	**0.253**	0.245	0.160
chair	**0.397**	**0.397**	0.312
cow	0.158	**0.177**	0.117
diningtable	**0.263**	0.227	0.245
dog	**0.261**	0.179	0.152
horse	**0.495**	0.450	0.437
motorbike	**0.295**	0.284	0.207
person	**0.630**	0.614	0.237
pottedplant	0.102	**0.116**	0.065
sheep	**0.184**	0.175	0.144
sofa	**0.246**	0.211	0.162
train	**0.399**	0.385	0.397
tvmonitor	**0.272**	0.257	0.230
Moyenne	**0.301**	0.292	0.234

TABLE 4.6 – *Resultats en MAP obtenus sur l'échantillon test de PascalVOC'07 avec une couche de noyau gaussien.*

contraintes moyennées semblent pertinentes. Cependant, la contrainte de classement n'aide pas lors du processus d'apprentissage : la version classique de P-MinCq montre les meilleurs résultats et est statistiquement comparable à P-MinCq$_{PWav}$ (une p-valeur de 0.2574 avec un test de Student). En fait, le compromis entre diversité et classement est ici difficile à mettre en œuvre puisque les neuf votants ne sont probablement pas assez expressifs. D'une part, les contraintes d'ordre se retrouvent plus difficiles à satisfaire et, d'autre part, la diversité des votants varie peu.

L'ajout d'une couche de noyaux permet alors d'augmenter cette expressivité. En effet, la table 4.6, montrant que les méthodes basées sur P-MinCq produisent les meilleurs résultats par rapport au classifieur-SVM, confirme que la diversité des votants est correctement modélisée par P-MinCq. De plus, P-MinCq$^{rbf}_{PWav}$ est cette fois-ci significativement le meilleur : un test de Student renvoie une p-valeur de 0.0003 lorsque l'on compare P-MinCq$^{rbf}_{PWav}$ et le classifieur-SVM ; la p-valeur est de 0.0038 lorsqu'on le compare avec P-MinCq^{rbf}. Ainsi, la contrainte moyennée est un bon compromis entre obtenir un bon MAP et garder un temps de résolution raisonnable. Notons que nous n'avons pas reporté les résultats obtenus avec P-MinCq$^{rbf}_{PW}$ puisque dans ce contexte, le temps d'exécution est beaucoup plus élevé avec une performance faible. En fait, la méthode par paires implique un trop grand nombre de variables, pénalisant alors la résolution de P-MinCq$^{rbf}_{PW}$.

Finalement, il s'avère qu'au moins une méthode de type P-MinCq est la meilleure pour chaque concept. De plus, un test de Student renvoie une p-valeur proche de 0 lorsque l'on compare P-MinCq$^{rbf}_{PWav}$ avec les méthodes n'utilisant pas la couche supplémentaire de noyaux. Ce résultat justifie que cette approche est significativement meilleure.

4.3.4 Conclusion

Les expériences menées dans cette section ont démontré que P-MinCq offre une alternative théoriquement fondée pour le problème de la fusion de classifieurs en prenant en considération la diversité des votants. Dans le cadre particulier de l'indexation sémantique de documents multimédia, la diversité des votants est induite par la variabilité des descripteurs ou par la variabilité de la première couche de classifieurs. De plus, une contrainte d'ordre permet d'améliorer les performances en terme de MAP. Contrairement à la section 4.2, nous n'avons pas pu définir de contrainte de π-alignement adéquate à ce problème d'apprentissage multimodal. Une de nos perspectives est ainsi de définir une telle contrainte en étudiant les travaux déjà réalisés sur la notion d'apprentissage de prior en théorie PAC-Bayésienne. Nous pourrions, par exemple, apprendre un contrainte *a priori* à partir d'une fraction de l'échantillon d'apprentissage à l'image de [Ambroladze *et al.*, 2006, Parrado-Hernández *et al.*, 2012].

4.4 Synthèse

Dans ce chapitre, nous avons répondu aux limitations de l'algorithme MinCq. Nous en avons proposé une généralisation, nommée P-MinCq, permettant l'introduction d'une connaissance *a priori* lors du processus d'apprentissage. Tandis que la contrainte de quasi-uniformité imposée par MinCq est non informative, la connaissance prise en compte par P-MinCq prend la forme d'une distribution π-alignée. D'une part, notre approche n'implique aucune perte d'expressivité et, d'autre part, nous en avons démontré des garanties en généralisation lorsque les votants dépendent des données. Nous avons réalisé deux instanciations de P-MinCq. Une première pour combiner des classifieurs de type k-PPV, mettant en évidence l'intérêt du π-alignement lors du processus d'apprentissage. Une seconde dans la situation particulière de la fusion de classifieurs en multimédia, montrant l'intérêt d'une telle approche pour apprendre un vote de majorité pondéré tout en prenant en compte la diversité des classifieurs.

Ces résultats prometteurs ouvrent des perspectives concernant la définition de π-alignements pour différents types de classifieurs, mais aussi pour mettre au point des méthodes visant à estimer π pour une tâche particulière. En outre, dans le cadre spécifique des k-PPV, il serait intéressant de combiner P-MinCq avec la méthode d'apprentissage de distance χ^2 proposée très récemment [Kedem *et al.*, 2012].

Dans ce chapitre, nous nous sommes focalisés sur la classification binaire, ce qui est une limitation importante de cette contribution. Dans le chapitre suivant, nous abordons la problématique de la classification multiclasse. L'extension de ces travaux à une telle situation n'est pas aisée. Ainsi, dans un premier temps, nous proposons une analyse PAC-Bayésienne sur le classifieur de Gibbs, puis, dans un second temps nous généralisons la C-borne au multiclasse.

Théorie PAC-Bayésienne et classification multiclasse

5

Alors que les contributions du chapitre précédent se placent dans le cadre de la classification binaire, nous présentons dans ce chapitre deux travaux en théorie PAC-Bayésienne pour la classification multiclasse lorsque le nombre de classes Q est fini et supérieur à 2.

Dans un premier temps en section 5.2 nous démontrons une borne PAC-Bayésienne sur le risque du classifieur de Gibbs, qui, nous le rappelons, correspond à la moyenne selon une distribution ρ des risques des classifieurs d'un ensemble $\mathcal{H}$. L'originalité et l'intérêt de ce résultat est de considérer une mesure de risque basée sur la matrice de confusion des classifieurs. En effet, en classification multiclasse, la matrice de confusion offre un outil plus approprié et plus riche que le simple taux d'erreur mesuré par la fonction de perte $0-1$. En fait, elle permet l'étude explicite des différentes probabilités d'erreur, autrement dit, les probabilités qu'un exemple de classe y soit classé en une classe différente $y' \neq y$. Cette notion est effectivement cruciale dans de nombreuses applications. Dans le cas de la classification binaire, il est souvent intéressant de distinguer les faux-positifs (test positif à tort) des faux-négatifs (test négatif à tort). Par exemple, en médecine, il est important de différencier "faux-malade" et "faux-sain" : il vaut mieux traiter à tort un sujet en réalité sain que de ne pas traiter un sujet malade. Concrètement, nous faisons appel à un résultat récent sur les inégalités de concentrations valables sur des sommes de matrices aléatoires [Tropp, 2011] pour dériver deux bornes montrant que la norme opérateur de la matrice de confusion réelle du classifieur de Gibbs est bornée par : (i) la norme opérateur de son estimation empirique,

(ii) un terme dépendant de la quantité d'exemples dans chacune des classes et (iii) la KL-divergence entre les distributions prior et posterior. À notre connaissance, il s'agit de la première borne PAC-Bayésienne pour la matrice de confusion. Notons que des garanties en généralisation sur la norme opérateur de la matrice de confusion ont été récemment démontrées dans d'autres cadres tels que celui de la stabilité uniforme [Machart, 2012] ou de l'apprentissage en ligne [Ralaivola, 2012].
Dans un second temps, en section 5.3, nous étudions les relations qui existent entre l'erreur du classifieur de Gibbs et l'erreur du vote de majorité. Tout d'abord, nous montrons que la relation directe repose sur un facteur égal au nombre de classes. Puis, nous généralisons la C-borne du théorème 3.3 du chapitre 3. Nous en donnons plusieurs formulations reposant sur des notions toutes équivalentes en classification binaire. Nous rappelons que la C-borne est au cœur des algorithmes MinCq (section 3.4, chapitre 3) et P-MinCq du chapitre précédent. Cependant, à ce jour, il s'avère plus complexe de dériver un algorithme performant. Nous discutons de cet aspect à la fin de ce chapitre.

Les travaux sur la matrice de confusion on été publiés à ICML 2012 [Morvant *et al.*, 2012c]. La généralisation de la C-borne a donné lieu à une communication scientifique non publiée dans le workshop WiML [1] 2012.

5.1 Le cadre multiclasse considéré et quelques notations

Nous étudions, ici, des tâches de classification multiclasse où $X \subseteq \mathbb{R}^d$, de dimension finie d, est l'espace d'entrée et $Y = \{1, \ldots, Q\}$ est l'espace de sortie avec $Q > 2$ le nombre de classes fini. L'ensemble d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ est constitué de m exemples tirés *i.i.d.* selon un domaine P défini sur $X \times Y$, tel que $m \geq Q$ et $m_y \geq 1$ pour toutes les classes y de Y, où m_y est le nombre d'exemples de vraie classe y. $\mathcal{H}$ est une famille de classifieurs multiclasse de X vers Y. Nous rappelons qu'étant donnés une distribution prior π sur $\mathcal{H}$ et un échantillon S, le processus d'apprentissage PAC-Bayésien vise à trouver la distribution posterior ρ amenant à minimiser l'erreur réelle du classifieur stochastique de Gibbs $G_\rho(\cdot)$ ou celle du vote de majorité ρ-pondéré $B_\rho(\cdot)$. Pour rappel, le classifieur de Gibbs $G_\rho(\cdot)$ étiquette un exemple $\mathbf{x}$ en tirant aléatoirement selon ρ un votant h dans $\mathcal{H}$, puis en renvoyant la valeur de $h(\mathbf{x})$. Le vote de majorité ρ-pondéré $B_\rho(\cdot)$ est, quant à lui, défini en classification multiclasse par :

$$B_\rho(\mathbf{x}) = \operatorname*{argmax}_{c \in Y} \left[\mathop{\mathbf{E}}_{h \sim \rho} \mathbf{I}(h(\mathbf{x}) = c) \right]. \tag{5.1}$$

Cependant, dans de nombreuses situations — par exemple lors d'un déséquilibre entre les classes — mesurer la qualité d'un classifieur uniquement en comptabilisant ses erreurs n'est pas judicieux. En effet, supposons qu'une classe p soit majoritairement présente, c'est-à-dire que sa probabilité d'apparition soit de $1 - \epsilon$ avec $\epsilon > 0$ petit, alors le classifieur $h_{maj}(\cdot)$, qui renvoie la classe p pour tous les exemples $\mathbf{x}$, montrera un taux d'erreur d'au plus ϵ, alors qu'il commet toujours une erreur pour les exemples de

1. *Women in Machine Learning*, `http://wimlworkshop.org/`.

classes $q \neq p$.

Pour contourner ce problème, nous faisons appel à la matrice de confusion du classifieur en guise de mesure du risque plus fine. En particulier, nous étudions la matrice de confusion construite à partir de sa définition classique qui se base sur les probabilités conditionnelles : il est, en effet, important de s'affranchir des effets de la diversité des classes représentées. Concrètement, pour un classifieur h issu de $\mathcal{H}$ et un échantillon d'apprentissage S, la matrice de confusion empirique $\mathbf{D}_S^h = (\hat{d}_{pq})_{1\leq p,q\leq Q}$ associée à h est définie par :

$$\forall (p,q) \in Y^2,\ \hat{d}_{pq} = \sum_{i=1}^{m} \frac{1}{m_{y_i}} \mathbf{I}(h(\mathbf{x}_i) = q)\mathbf{I}(y_i = p).$$

La matrice de confusion réelle $\mathbf{D}_P^h = (d_{pq})_{1\leq p,q\leq Q}$ associée à h est :

$$\begin{aligned}\forall (p,q) \in Y^2,\ d_{pq} &= \mathop{\mathbf{E}}_{\mathbf{x}|y=p} \mathbf{I}\big(h(\mathbf{x}) = q\big) \\ &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} (h(\mathbf{x}) = q | y = p).\end{aligned}$$

Si le classifieur h ne commet aucune erreur sur l'échantillon S, alors tous les éléments de la matrice de confusion sont nuls, à l'exception des éléments diagonaux tous égaux à 1 puisqu'ils correspondent aux exemples correctement classés. Ainsi, plus il y a d'éléments non nuls en dehors de la diagonale, plus le classifieur est enclin à commettre des erreurs. Minimiser le nombre d'erreurs revient donc à trouver un classifieur dont la matrice de confusion montre le plus possible d'éléments faibles en dehors de la diagonale.

Puisque la diagonale porte l'information des probabilités conditionnelles des prédictions "correctes", nous proposons d'annuler tous ces éléments. Les éléments non nuls de cette nouvelle matrice de confusion correspondent alors, uniquement, aux exemples mal classés par h. Nous la définissons comme suit.

Définition 5.1 (Matrice de confusion empirique et réelle) *Soit $Y = \{1, \ldots, Q\}$ l'ensemble des classes possibles. Soit P un domaine sur $X \times Y$. Soit $S \sim (P)^m$ un échantillon d'apprentissage de m exemples. Soit $\mathcal{H}$ un ensemble de classifieurs multiclasse de X vers Y. Alors, pour tout classifieur h issu de $\mathcal{H}$, nous définissons les matrices de confusion empirique et réelle de h respectivement par $\mathbf{C}_S^h = (\hat{c}_{pq})_{1\leq p,q\leq Q}$ et $\mathbf{C}_P^h = (c_{pq})_{1\leq p,q\leq Q}$ telles que :*

$$\forall (p,q) \in Y^2,\ \hat{c}_{pq} = \begin{cases} 0 & si\ q = p \\ \hat{d}_{pq} & sinon, \end{cases}$$

$$\forall (p,q) \in Y^2,\ c_{pq} = \begin{cases} 0 & si\ q = p \\ d_{pq} = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} (h(\mathbf{x}) = q | p = y) & sinon. \end{cases}$$

Par abus de langage, c'est cette définition que nous nommons "matrice de confusion". Notons qu'elle n'implique aucune perte d'expressivité car l'information portée par la

diagonale de la matrice de confusion classique est redondante :

$$\begin{aligned}
&\forall p \in Y,\ \sum_{q \in Y} d_{pq} = 1 \\
\Leftrightarrow\ &\forall p \in Y,\ d_{pp} = 1 - \sum_{q \in Y,\ q \neq p} d_{pq} \\
\Leftrightarrow\ &\forall p \in Y,\ d_{pp} = 1 - \sum_{q \in Y} c_{pq}.
\end{aligned}$$

De plus, il est facile de montrer que si $\mathbf{D}_Y = \Big(\mathbf{Pr}\,(y=1), \ldots, \mathbf{Pr}\,(y=Q)\Big)^\top$ est le vecteur des probabilités *a priori* des classes, alors le taux d'erreur est égal à :

$$\mathbf{R}_P(h) = \left\| {\mathbf{D}_Y}^\top \mathbf{C}_P^h \right\|_1 .$$

Ainsi, à la lumière de l'information supplémentaire apportée par un échantillon représentatif, le taux d'erreur peut être retrouvé à partir de la matrice de confusion (l'inverse étant impossible).

Si h classe correctement tous les exemples issus de S, alors sa matrice de confusion empirique $\mathbf{C}_S^h$ correspond simplement à la matrice nulle $\mathbf{0}$. De même, si h est un classifieur parfait sur le domaine P, alors sa matrice de confusion réelle vaut $\mathbf{0}$. Contrôler la matrice de confusion d'un classifieur apparaît donc être une stratégie adéquate pour contrôler l'erreur du classifieur de Gibbs. Plus précisément, une solution consiste à chercher la matrice de confusion la plus "petite" possible, où "petite" signifie la plus proche possible de $\mathbf{0}$. Ici, nous choisissons comme mesure de distance à $\mathbf{0}$ la norme opérateur, aussi appelée norme spectrale. Cette norme retourne la plus grande valeur singulière de son argument et est définie par :

$$\begin{aligned}
\|\mathbf{C}\|_{op} &= \max\{\lambda_{\max}(\mathbf{C}), -\lambda_{\min}(\mathbf{C})\} \\
&= \max_{\mathbf{v} \neq \mathbf{0}} \frac{\|\mathbf{C}\mathbf{v}\|_2}{\|\mathbf{v}\|_2},
\end{aligned} \tag{5.2}$$

où $\lambda_{\max}(\mathbf{C})$ et $\lambda_{\min}(\mathbf{C})$ sont respectivement les valeurs singulières maximale et minimale de la matrice $\mathbf{C}$. Le choix de la norme opérateur n'est pas anodin. D'une part, nos travaux tirent parti d'un des résultats de [Tropp, 2011] portant sur la norme opérateur de matrices aléatoires. D'autre part, on peut prouver la relation suivante :

$$\begin{aligned}
\mathbf{R}_P(h) &= \left\| {\mathbf{D}_Y}^\top \mathbf{C}_P^h \right\|_1 \\
&= \mathbf{D}_Y{}^\top \mathbf{C}_P^h\, \mathbf{Id} && \text{où } \mathbf{Id} \text{ est le vecteur identité,} \\
&\leq \sqrt{Q} \left\| {\mathbf{D}_Y}^\top \mathbf{C}_P^h \right\|_2 && \text{d'après l'inégalité de Cauchy-Swartz,} \\
&= \sqrt{Q} \left\| {\mathbf{C}_P^h}^\top \mathbf{D}_Y \right\|_2 \\
&\leq \sqrt{Q} \left\| {\mathbf{C}_P^h}^\top \right\|_{op} \|\mathbf{D}_Y\|_2 && \text{d'après la définition de } \|\cdot\|_{op} \text{ de l'équation (5.2),} \\
&\leq \sqrt{Q} \left\| {\mathbf{C}_P^h}^\top \right\|_{op} && \text{puisque } \|\mathbf{D}_Y\|_2 \leq 1, \\
&= \sqrt{Q} \left\| \mathbf{C}_P^h \right\|_{op} .
\end{aligned}$$

Ainsi, la minimisation de la norme opérateur $\left\|\mathbf{C}_P^h\right\|_{op}$ permet de minimiser l'erreur.

Finalement, nous précisons quelques propriétés de la norme opérateur. Tout d'abord, $\|\cdot\|_{op}$ est une norme régulière, pour toute matrice $\mathbf{C}$, on a donc :

$$\forall a \in \mathbb{R},\ \|a\mathbf{C}\|_{op} = |a|\,\|\mathbf{C}\|_{op}\,. \tag{5.3}$$

De plus, étant données les matrices de même dimensions $\mathbf{A}$ et $\mathbf{B}$ composées d'éléments positifs ou nuls et telles que $0 \leq \mathbf{A} \leq \mathbf{B}$ (éléments par éléments), on a :

$$0 \leq \mathbf{A} \leq \mathbf{B} \implies \|\mathbf{A}\|_{op} \leq \|\mathbf{B}\|_{op}\,. \tag{5.4}$$

À partir de ces éléments, nous démontrons tout d'abord une borne PAC-Bayésienne pour la matrice de confusion du classifieur de Gibbs. Nous étudions ensuite le lien entre le classifieur de Gibbs et le vote de majorité ρ-pondéré en section 5.3.

5.2 Borne PAC-Bayésienne sur la confusion du classifieur de Gibbs

5.2.1 La borne en généralisation

Le résultat principal de cette section est une borne PAC-Bayésienne en généralisation sur la matrice de confusion associée au classifieur stochastique de Gibbs $G_\rho(\cdot)$ dans le contexte de la prédiction multiclasse décrit précédemment. Dans ce cas, les matrices de confusion réelle et empirique associées à $G_\rho(\cdot)$ correspondent à une espérance selon la distribution posterior ρ et sont respectivement notées :

$$\mathbf{C}_P^{G_\rho} = \mathop{\mathbf{E}}_{h\sim\rho}\ \mathop{\mathbf{E}}_{S\sim(P)^m}\ \mathbf{C}_S^h \qquad \text{et} \qquad \mathbf{C}_S^{G_\rho} = \mathop{\mathbf{E}}_{h\sim\rho}\ \mathbf{C}_S^h.$$

Étant donnés un votant h tiré aléatoirement selon le posterior ρ et un échantillon $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (P)^m$, notre objectif est de majorer l'écart entre $\mathbf{C}_P^{G_\rho}$ et $\mathbf{C}_S^{G_\rho}$ qui sont vues comme les mesures des risques réel et empirique du classifieur de Gibbs. Notre résultat principal est énoncé dans le théorème suivant.

Théorème 5.1 *Soit $X \subseteq \mathbb{R}^d$ l'espace d'entrée, soit $Y = \{1, \ldots, Q\}$ l'ensemble de classes. Soit P un domaine sur $X \times Y$ et soit $\mathcal{H}$ une famille de classifieurs multiclasse de X vers Y. Soit S un échantillon d'apprentissage constitué de m exemples tirés i.i.d. selon P et tel que pour toute classe y de Y, on ait $m_y > 8Q$. Alors pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0, 1]$, avec une probabilité d'au moins $1 - \delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$ et pour toute distribution ρ sur $\mathcal{H}$, on a :*

$$\left\|\mathbf{C}_S^{G_\rho} - \mathbf{C}_P^{G_\rho}\right\|_{op} \leq \sqrt{\frac{8Q}{m_- - 8Q}\left[\mathrm{KL}(\rho||\pi) + \ln\left(\frac{m_-}{4\delta}\right)\right]},$$

où $m_- = \min_{y\in Y} m_y$ est le nombre minimal d'exemples de S de même classe.

Démonstration. Reportée en section 5.2.2. □

Notons que pour toute classe y de l'ensemble Y, il est nécessaire de vérifier : $m_y > 8Q$. Cette restriction n'est pas trop forte puisqu'elle est linéaire en le nombre de classes Q.

Finalement, on peut réécrire le théorème 5.1 de la manière suivante :

Corollaire 5.1 *Sous les hypothèses du théorème 5.1, on a :*

$$\left\|\mathbf{C}_P^{G_\rho}\right\|_{op} \leq \left\|\mathbf{C}_S^{G_\rho}\right\|_{op} + \sqrt{\frac{8Q}{m_- - 8Q}\left[\mathrm{KL}(\rho||\pi) + \ln\left(\frac{m_-}{4\delta}\right)\right]}.$$

Démonstration. On applique l'inégalité triangulaire $|\|\mathbf{A}\| - \|\mathbf{B}\|| \leq \|\mathbf{A} - \mathbf{B}\|$ au théorème 5.1. □

En se fixant une distribution prior π sur $\mathcal{H}$, le théorème 5.1 et le corollaire 5.1 donnent une borne PAC-Bayésienne sur l'estimation, via la norme opérateur, de la matrice de confusion réelle du classifieur de Gibbs pour toutes les distributions prior et posterior possibles. Pour une tâche donnée, le nombre de classes est constant, le risque réel est donc majoré par le risque empirique du classifieur de Gibbs, la KL-divergence entre ρ et π et un terme relié aux nombres d'exemples d'apprentissage. Ce dernier dépend en fait de m_-, la quantité minimale d'exemples qui appartiennent à la même classe : plus m_- est élevé, plus la borne sera précise. Cependant, si une classe est très minoritairement présente, la borne sera imprécise. Notons finalement que ce résultat varie en $O(1/\sqrt{m_-})$ qui est un taux de convergence typique pour une borne ne contenant pas d'information du second ordre.

5.2.2 Démonstration du résultat

Cette section énonce la preuve du théorème 5.1. Tout d'abord, nous introduisons une inégalité de concentration généralisant l'inégalité de Hoeffding à la somme de matrices carrées aléatoires. Nous en déduisons ensuite notre borne en généralisation PAC-Bayésienne en suivant le principe de preuve "en trois étapes", qui se base sur l'inégalité de Hoeffding, du théorème PAC-Bayes binaire de Langford-Seeger [Seeger, 2002, McAllester, 2003, Langford, 2005] (le corollaire 3.1 vu en chapitre 3).

Inégalité de concentration pour la matrice de confusion

Le résultat principal de ce chapitre se dérive à l'aide de l'inégalité de concentration suivante sur la somme de matrices auto-adjointes démontrée par [Tropp, 2011]. Cette inégalité généralise l'inégalité de Hoeffding aux matrices aléatoires auto-adjointes [2] (ou hermitienne).

Théorème 5.2 ([Tropp, 2011]) *Considérons une séquence finie $\{\mathbf{M}_i\}$ de matrices auto-adjointes, aléatoires et indépendantes de dimension Q. Soit $\{\mathbf{A}_i\}$ une séquence de matrices auto-adjointes.*

2. Une matrice auto-adjointe (ou hermitienne) est une matrice carrée qui est égale à la matrice tansposée de la matrice conjuguée.

Supposons que chaque matrice carrée aléatoire vérifie presque sûrement :

$$\mathbf{E}\,\mathbf{M}_i = \mathbf{0} \quad et \quad \mathbf{M}_i^2 \preccurlyeq \mathbf{A}_i^2.$$

Alors pour tout $\epsilon \geq 0$, on a :

$$\mathbf{Pr}\left\{\lambda_{\max}\Big(\sum_i \mathbf{M}_i\Big) \geq \epsilon\right\} \leq Q \exp\left(\frac{-\epsilon^2}{8\sigma^2}\right),$$

où $\sigma^2 = \left\|\sum_i \mathbf{A}_i^2\right\|$ et $\preccurlyeq$ est l'ordre semi-défini sur les matrices auto-adjointes.

Sachant qu'une matrice à valeurs réelles est auto-adjointe si et seulement si elle est symétrique, la matrice de confusion qui, elle, l'est rarement n'est donc en général pas auto-adjointe. Pour adapter ce résultat à notre situation particulière, nous faisons appel à la méthode de dilatation de matrices [Paulsen, 2002] pour "transformer" une matrice carrée en une matrice auto-adjointe, tout en gardant l'information spectrale importante pour la norme opérateur.

Définition 5.2 (Dilatation [Paulsen, 2002]) *La dilatation $\mathcal{D}(\mathbf{M})$ d'une matrice $\mathbf{M}$ carrée réelle de dimension Q est la matrice de dilatation carrée auto-adjointe $\mathcal{D}(\mathbf{M})$ de dimension $2Q$ définie par :*

$$\mathcal{D}(\mathbf{M}) = \begin{pmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M}^\top & \mathbf{0} \end{pmatrix}. \tag{5.5}$$

De plus, la dilatation préserve l'information spectrale :

$$\begin{aligned} \lambda_{\max}\left(\mathcal{D}(\mathbf{M})\right) &= \left\|\mathcal{D}(\mathbf{M})\right\|_{op} \\ &= \left\|\mathbf{M}\right\|_{op}. \end{aligned} \tag{5.6}$$

À l'aide de cette définition, nous prouvons l'inégalité de concentration suivante.

Corollaire 5.2 *Considérons une séquence finie $\{\mathbf{M}_i\}$ de matrices carrées aléatoires et indépendantes de dimension Q. Soit $\{a_i\}$ une séquence de réels fixés. Supposons que chaque matrice aléatoire vérifie presque sûrement :*

$$\mathbf{E}\,\mathbf{M}_i = \mathbf{0} \quad et \quad \left\|\mathbf{M}_i\right\|_{op} \leq a_i.$$

Alors, pour tout $\epsilon \geq 0$, on a :

$$\mathbf{Pr}\left\{\left\|\sum_i \mathbf{M}_i\right\|_{op} \geq \epsilon\right\} \leq 2Q \exp\left(\frac{-\epsilon^2}{8\sigma^2}\right), \tag{5.7}$$

où $\sigma^2 = \sum_i a_i^2$.

Démonstration. Pour prouver ce corollaire, il faut partir de la définition 5.2 pour vérifier les hypothèses du théorème 5.2 afin de l'appliquer.
Soit $\{\mathbf{M}_i\}$ une séquence de matrices carrées, aléatoires et indépendantes de dimension Q telles que : $\mathbf{E}\,\mathbf{M}_i = \mathbf{0}$. Soit $\{a_i\}$ une séquence de réels fixés tels que : $\left\|\mathbf{M}_i\right\|_{op} \leq a_i$.
Nous considérons les matrices de dilatation $\{\mathcal{D}(\mathbf{M}_i)\}$ de dimension $2Q$ associées

aux matrices $\{\mathbf{M}_i\}$. Par définition, la dilatation est un opérateur linéaire et donc : $\mathbf{E}\ \mathcal{D}(\mathbf{M}_i) = \mathbf{0}$. En outre, d'après l'équation 5.6, on a :

$$\begin{aligned}\left\|\sum_i \mathbf{M}_i\right\|_{op} &= \lambda_{\max}\left(\mathcal{D}\left(\sum_i \mathbf{M}_i\right)\right)\\ &= \lambda_{\max}\left(\sum_i \mathcal{D}(\mathbf{M}_i)\right).\end{aligned}$$

De plus :

$$\begin{aligned}\|\mathbf{M}_i\|_{op} &= \|\mathcal{D}(\mathbf{M}_i)\|\\ &= \lambda_{\max}\left(\mathcal{D}(\mathbf{M}_i)\right)\\ &\leq a_i.\end{aligned}$$

Il ne reste plus qu'à vérifier l'hypothèse : $\mathcal{D}(\mathbf{M}_i)^2 \preccurlyeq \mathbf{A}_i^2$. Pour ce faire, nous devons définir et fixer une séquence de matrices auto-adjointes $\{\mathbf{A}_i\}$ de dimension $2Q$. Il suffit de construire la matrice diagonale dont tous les éléments de la diagonale valent $\lambda_{\max}\big(\mathcal{D}(\mathbf{M}_i)\big)$:

$$\lambda_{\max}\big(\mathcal{D}(\mathbf{M}_i)\big)\mathbf{Id}_{2Q}.$$

Ceci assure :

$$\mathcal{D}(\mathbf{M}_i)^2 \preccurlyeq \left(\lambda_{\max}\left(\mathcal{D}(\mathbf{M}_i)\right)\mathbf{Id}_{2Q}\right)^2.$$

Plus précisément, pour tout i on a :

$$\lambda_{\max}\big(\mathcal{D}(\mathbf{M}_i)\big) \leq a_i.$$

Nous fixons $\mathbf{A}_i$ la matrice diagonale de coefficients tous égaux à a_i telle que :

$$\mathbf{A}_i = a_i\mathbf{Id}_{2Q},$$

avec :

$$\begin{aligned}\left\|\sum_i \mathbf{A}_i^2\right\|_{op} &= \sum_i a_i^2\\ &= \sigma^2.\end{aligned}$$

Finalement, on applique le théorème 5.2 pour obtenir l'inégalité de concentration (5.7). □

Pour appliquer directement ce corollaire à la matrice de confusion, nous devons la réécrire comme la somme des matrices de confusion mesurées indépendamment sur chaque exemple. Étant donné un exemple $(\mathbf{x}_i, y_i)$ de l'échantillon S, nous définissons sa matrice de confusion $\mathbf{C}_i^h = (\hat{c}_{pq}(i))_{1\leq p,q\leq Q}$ par :

$$\forall (p,q) \in Y^2,\ \hat{c}_{pq}(i) = \begin{cases} 0 & \text{si } q = p\\ \dfrac{1}{m_{y_i}}\mathbf{I}(h(\mathbf{x}_i) = q)\mathbf{I}(y_i = p) & \text{sinon,}\end{cases}$$

où m_{y_i} est le nombre d'exemples de S de classe y_i. Étant donné un exemple $(\mathbf{x}_i, y_i)$, la matrice de confusion définie sur $(\mathbf{x}_i, y_i)$ contient au plus un élément non nul lorsque $h(\mathbf{x}_i)$ ne renvoie pas y_i. De la même manière, lorsque $h(\mathbf{x}_i)$ renvoie y_i alors la matrice de confusion définie sur l'exemple est égale à $\mathbf{0}$. Ainsi, pour tout échantillon $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ composé de m éléments *i.i.d.* selon P et pour tout classifieur h issu de $\mathcal{H}$, la matrice de confusion est équivalente à :

$$\mathbf{C}_S^h = \sum_{i=1}^m \mathbf{C}_i^h.$$

En outre, on considère les matrices carrées aléatoires $\mathbf{C'}_i^h = (\hat{c}'_{pq}(i))_{1\leq p,q\leq Q}$ définies par :

$$\forall (p,q) \in Y^2,$$
$$\hat{c}'_{pq}(i) = \begin{cases} 0 & \text{si } \hat{c}_{pq}(i) = 0 \\ \dfrac{1}{m_{y_i}} \left(\mathbf{I}(h(\mathbf{x}_i) = q)\mathbf{I}(y_i = p) - \underset{\substack{S=\{(\mathbf{x}_j,y_j)\}_{j=1}^m \\ \sim (P)^m}}{\mathbf{E}} \dfrac{1}{m_{y_j}} \mathbf{I}(h(\mathbf{x}_j) = q)\mathbf{I}(y_j = p) \right) & \text{sinon.} \end{cases} \tag{5.8}$$

Le terme $\mathbf{E}_{S\sim(P)^m} \frac{1}{m_{y_j}} \mathbf{I}(h(\mathbf{x}_j) = q)\mathbf{I}(y_j = p)$, lorsque $\hat{c}_{pq}(i) \neq 0$, est équivalent à[3] l'espérance sur $S \sim (P)^m$ des éléments $\hat{c}_{pq}$, tels que $p = y_i$ et $q = h(\mathbf{x}_i)$, issus des matrices empiriques $\mathbf{C}_S^h$. L'équation (5.8) se réécrit alors :

$$\forall (p,q) \in Y^2, \hat{c}'_{pq}(i) = \begin{cases} 0 & \text{si } \hat{c}_{pq}(i) = 0 \\ \hat{c}_{pq}(i) - \dfrac{1}{m_{y_i}} \underset{S\sim(P)^m}{\mathbf{E}} \hat{c}_{pq} & \text{sinon.} \end{cases} \tag{5.9}$$

Par souci de lisibilité, étant donné un exemple $(\mathbf{x}_i, y_i)$ et pour tout échantillon $S \sim (P)^m$, nous notons $\mathbf{C}_{S|i}^h$ la matrice contenant au plus un élément non nul de coordonnées (p,q) qui vaut $\hat{c}_{pq}$ avec $p = y_i$ et $q = h(\mathbf{x}_i)$. On obtient alors la formulation de $\mathbf{C'}_i^h$ suivante :

$$\mathbf{C'}_i^h = \mathbf{C}_i^h - \frac{1}{m_{y_i}} \underset{S\sim(P)^m}{\mathbf{E}} \mathbf{C}_{S|i}^h ,$$

Trivialement, cette définition nous permet de vérifier : $\mathbf{E}\ \mathbf{C'}_i^h = 0$.
Il ne reste plus qu'à définir une valeur adéquate pour le réel a_i associé à chaque matrice $\mathbf{C'}_i^h$. Soit $\lambda_{\max_i}(\mathbf{C'}_i^h)$ la valeur singulière maximale de la matrice $\mathbf{C'}_i^h$. Il est facile de vérifier que $\lambda_{\max_i}(\mathbf{C'}_i^h) \leq \frac{1}{m_{y_i}}$. Ainsi, nous posons :

$$\forall i \in \{1, \ldots, m\},\ a_i = \frac{1}{m_{y_i}}. \tag{5.10}$$

Finalement, les notations, précédemment introduites, nous permettent d'appliquer le corollaire 5.2 pour obtenir l'inégalité de concentration suivante :

$$\mathbf{Pr}\left\{ \left\| \sum_{i=1}^m \mathbf{C'}_i^h \right\|_{op} \geq \epsilon \right\} \leq 2Q \exp\left(\frac{-\epsilon^2}{8\sigma^2} \right). \tag{5.11}$$

3. L'idée est la suivante : on considère uniquement les éléments dont les coordonnées sont associées à l'élément non nul de $\mathbf{C}_i^h$.

Cette inégalité de concentration (5.11) nous permet de démontrer notre théorème 5.1 en suivant le principe de preuve "en trois étapes" [Seeger, 2002, McAllester, 2003, Langford, 2005].

La preuve "en trois étapes"

Étant donnée notre inégalité de concentration (5.11), nous allons appliquer la méthode de preuve proposée par [Seeger, 2002, McAllester, 2003, Langford, 2005].

Étape 1 Nous démontrons tout d'abord le lemme suivant :

Lemme 5.1 *Soit Q la dimension de $\mathbf{C}_S^h$ et $\mathbf{C'}_i^h = \mathbf{C}_i^h - \frac{1}{m_{y_i}} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathbf{C}_{S|i}^h$ définies dans l'équation* (5.9). *Alors pour tout $\delta \in (0,1]$, on a :*

$$\mathop{\mathbf{Pr}}_{S\sim(P)^m} \left\{ \mathop{\mathbf{E}}_{h\sim\pi} \left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2} \left\| \sum_{i=1}^{m} \mathbf{C'}_i^h \right\|_{op}^2 \right) \right] \leq \frac{2Q}{8\sigma^2\delta} \right\} \geq 1-\delta$$

Démonstration. Par souci de lisibilité, nous posons :

$$\mathbf{C'}_S^h = \sum_{i=1}^{m} \mathbf{C'}_i^h.$$

Si Z est une variable aléatoire réelle telle que :

$$\mathbf{Pr}\ (Z \geq z) \leq k \exp\big[-ng(z)\big],$$

avec $g(z)$ ni négative, ni décroissante et k une constante positive, alors :

$$\mathbf{Pr}\ \big(\exp\big[(n-1)g(Z)\big] \geq \nu\big) \leq \min(1, k\nu^{-n/(n-1)}).$$

Appliqué à notre inégalité de concentration (5.11), en posant $g(z) = z^2$ (non négative), $z = \epsilon$, $n = \frac{1}{8\sigma^2}$ et $k = 2Q$, on obtient :

$$\mathbf{Pr}\ \left\{ \exp\left(\frac{1-8\sigma^2}{8\sigma^2} \left\| \mathbf{C'}_S^h \right\|_{op} \right) \geq \nu \right\} \leq \min\left(1, 2Q\nu^{-1/(1-8\sigma^2)}\right).$$

Comme le terme $\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C'}_S^h\|\right)$ est toujours strictement positif, on calcule son espérance :

$$\begin{aligned}
\mathbf{E}\ \left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2} \left\| \mathbf{C'}_S^h \right\|_{op} \right) \right] &= \int_0^\infty \mathbf{Pr}\ \left\{ \exp\left(\frac{1-8\sigma^2}{(8\sigma^2)} \left\| \mathbf{C'}_S^h \right\|_{op} \right) \geq \nu \right\} d\nu \\
&\leq 2Q + \int_1^\infty 2Q\nu^{-1/(1-8\sigma^2)} d\nu \\
&= 2Q - 2Q\frac{1-8\sigma^2}{8\sigma^2} \left[\nu^{-8\sigma^2/(1-8\sigma^2)} \right]_1^\infty \\
&= 2Q + 2Q\frac{1-8\sigma^2}{8\sigma^2} \\
&= \frac{2Q}{8\sigma^2}.
\end{aligned}$$

Donc, étant donné un classifieur h issu de $\mathcal{H}$, on a :

$$\mathop{\mathbf{E}}_{S\sim(P)^m}\left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\left\|\mathbf{C}'^h_S\right\|_{op}\right)\right]\leq\frac{2Q}{8\sigma^2}.$$

Alors si π est une distribution de probabilité sur l'ensemble $\mathcal{H}$, l'inégalité précédente implique :

$$\mathop{\mathbf{E}}_{S\sim(P)^m}\left[\mathop{\mathbf{E}}_{h\sim\pi}\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\left\|\mathbf{C}'^h_S\right\|_{op}\right)\right]\leq\frac{2Q}{8\sigma^2}.$$

On obtient le résultat du lemme en appliquant l'inégalité de Markov [4]. □

Étape 2 Cette étape de la preuve fait appel au lemme suivant.

Lemme 5.2 (Inégalité de Donsker-Varadhan [Donsker et Varadhan, 1975]) *Étant données la KL-divergence* $\mathrm{KL}(\rho\|\pi)$ *entre deux distributions* ρ *et* π*, et* $g(\cdot)$ *une fonction non négative, on a :*

$$\mathop{\mathbf{E}}_{a\sim\rho}\left[g(b)\right]\leq\mathrm{KL}(\rho\|\pi)+\ln\mathop{\mathbf{E}}_{b\sim\pi}\left[\exp(g(b))\right].$$

Démonstration. Voir [McAllester, 2003]. □

Rappelons que :

$$\mathbf{C}'^h_S=\sum_{i=1}^m\mathbf{C}'^h_i.$$

En posant $g(b)=\frac{1-8\sigma^2}{8\sigma^2}b^2$ et $b=\left\|\mathbf{C}'^h_S\right\|_{op}$, le lemme 5.2 implique :

$$\mathop{\mathbf{E}}_{h\sim\rho}\left[\frac{1-8\sigma^2}{8\sigma^2}\left\|\mathbf{C}'^h_S\right\|^2_{op}\right]\leq\mathrm{KL}(\rho\|\pi)+\ln\mathop{\mathbf{E}}_{h\sim\pi}\left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\left\|\mathbf{C}'^h_S\right\|^2_{op}\right)\right]. \tag{5.12}$$

Étape 3 La dernière étape consiste à appliquer à l'inégalité (5.12) au résultat du lemme 5.1. On a donc :

$$\mathop{\mathbf{E}}_{h\sim\rho}\left[\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C}'^h_S\|^2_{op}\right]\leq\mathrm{KL}(\rho\|\pi)+\ln\frac{2Q}{8\sigma^2\delta}.$$

Nous appliquons l'inégalité de Jensen [5] à la fonction convexe $g(\cdot)$. Alors avec une probabilité d'au moins $1-\delta$ sur le choix de $S\sim(P)^m$ et pour toute distribution ρ sur $\mathcal{H}$, on a :

$$\left(\mathop{\mathbf{E}}_{h\sim\rho}\left\|\mathbf{C}'^h_S\right\|_{op}\right)^2\leq\frac{8\sigma^2}{1-8\sigma^2}\left(\mathrm{KL}(\rho\|\pi)+\ln\frac{2Q}{8\sigma^2\delta}\right). \tag{5.13}$$

Puisque $\mathbf{C}'^h_S=\sum_{i=1}^m\left[\mathbf{C}^h_i-\frac{1}{m_{y_i}}\mathop{\mathbf{E}}_{S\sim(P)^m}\mathbf{C}^h_{S|i}\right]$, la borne (5.13) est similaire à celle énoncée dans le théorème 5.1. Nous présentons maintenant les simplifications menant à l'énoncé exact de notre borne en généralisation PAC-Bayésienne.

4. L'énoncé de l'inégalité de Markov est donné dans le théorème A.4 en annexe A.
5. L'énoncé de l'inégalité de Jensen est donné dans le théorème A.5 en annexe A.

Simplifications de la borne

Tout d'abord, nous calculons le paramètre de variance :

$$\sigma^2 = \sum_{i=1}^{m} a_i^2.$$

Rappelons que nous avions posé dans l'équation (5.10) :

$$\forall i \in \{1, \ldots, m\},\ a_i = \frac{1}{m_{y_i}},$$

où y_i est la classe du $i^{\text{ème}}$ exemple et m_{y_i} est le nombre d'exemples appartenant à la classe y_i. Ainsi :

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^{m} \frac{1}{m_{y_i}^2} \\ &= \sum_{y=1}^{Q} \sum_{i:y_i=y} \frac{1}{m_y^2} \\ &= \sum_{y=1}^{Q} \frac{1}{m_y}.\end{aligned}$$

Remarquons que le terme $\sum_{y=1}^{Q} \frac{1}{m_y}$ est croissant en fonction de Q, nous pouvons donc majorer σ^2 par :

$$\begin{aligned}\sigma^2 &= \sum_{y=1}^{Q} \frac{1}{m_y} \\ &\leq \frac{Q}{\min\limits_{y \in Y} m_y}.\end{aligned}$$

Soit $m_- = \min\limits_{y \in Y} m_y$, alors la borne (5.13) devient :

$$\left(\mathop{\mathbf{E}}_{h \sim \rho} \left[\left\| \mathbf{C'}_S^h \right\|_{op} \right] \right)^2 \leq \frac{8Q}{m_- - 8Q} \left(\mathrm{KL}(\rho \| \pi) + \ln \frac{m_-}{4\delta} \right).$$

Donc :

$$\mathop{\mathbf{E}}_{h \sim \rho} \left[\left\| \mathbf{C'}_S^h \right\|_{op} \right] \leq \sqrt{\frac{8Q}{m_- - 8Q} \left(\mathrm{KL}(\rho \| \pi) + \ln \frac{m_-}{4\delta} \right)}. \tag{5.14}$$

Il ne reste plus qu'à reformuler :

$$\mathbf{C'}_S^h = \sum_{i=1}^{m} \left[\mathbf{C}_i^h - \frac{1}{m_{y_i}} \mathop{\mathbf{E}}_{S \sim (P)^m} \mathbf{C}_{S|i}^h \right].$$

Rappelons que :

$$\mathbf{C}^{G_\rho} = \mathop{\mathbf{E}}_{h \sim \rho} \mathop{\mathbf{E}}_{S \sim (P)^m} \mathbf{C}_S^h \quad \text{et} \quad \mathbf{C}_S^{G_\rho} = \mathop{\mathbf{E}}_{h \sim \rho} \mathbf{C}_S^h.$$

On obtient :

$$\begin{aligned}
\mathop{\mathbf{E}}_{h\sim\rho}\left[\left\|\mathbf{C}'^{h}_{S}\right\|_{op}\right] &= \mathop{\mathbf{E}}_{h\sim\rho}\left[\left\|\sum_{i=1}^{m}\left[\mathbf{C}^{h}_{i}-\frac{1}{m_{y_i}}\mathop{\mathbf{E}}_{S\sim(P)^m}\mathbf{C}^{h}_{S|i}\right]\right\|_{op}\right]\\
&= \mathop{\mathbf{E}}_{h\sim\rho}\left[\left\|\sum_{i=1}^{m}\left[\mathbf{C}^{h}_{i}\right]-\sum_{i=1}^{m}\left[\frac{1}{m_{y_i}}\mathop{\mathbf{E}}_{S\sim(P)^m}\mathbf{C}^{h}_{S|i}\right]\right\|_{op}\right]\\
&= \mathop{\mathbf{E}}_{h\sim\rho}\left[\left\|\mathbf{C}^{h}_{S}-\mathop{\mathbf{E}}_{S\sim(P)^m}\mathbf{C}^{h}_{S}\right\|_{op}\right]\\
&\geq \left\|\mathop{\mathbf{E}}_{h\sim\rho}\left[\mathbf{C}^{h}_{S}-\mathop{\mathbf{E}}_{S\sim(P)^m}\mathbf{C}^{h}_{S}\right]\right\|_{op}\\
&= \left\|\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{C}^{h}_{S}-\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{S\sim(P)^m}\mathbf{C}^{h}_{S}\right\|_{op}\\
&= \left\|\mathbf{C}^{G_\rho}_{S}-\mathbf{C}^{G_\rho}\right\|_{op}. \qquad (5.15)
\end{aligned}$$

En substituant la partie gauche de l'inégalité (5.14) par le terme (5.15), nous obtenons la borne du théorème 5.1.

Ceci termine la preuve du théorème 5.1 : notre borne en généralisation pour le classifieur de Gibbs. Nous analysons maintenant les relations qui existent entre le classifieur de Gibbs et le vote de majorité ρ-pondéré. Tout d'abord, nous étudions une relation directe et linéaire, puis nous étendons la C-borne à la classification multiclasse.

5.3 Bornes sur le risque du vote de majorité ρ-pondéré

Dans la section précédente, nous nous sommes focalisés sur le risque du classifieur de Gibbs $G_\rho(\cdot)$. Comme précisé dans le chapitre 3, ce qui nous intéresse généralement est le risque du vote de majorité $B_\rho(\cdot)$, qui renvoie la classe majoritaire sous la mesure ρ. Nous allons donc maintenant étudier les relations qui existent entre ces deux classifieurs dans le cadre multiclasse.

5.3.1 Relation linéaire entre le classifieur de Gibbs et le vote de majorité

Notre borne multiclasse présentée dans le théorème 5.1 apporte des garanties théoriques pour le classifieur de Gibbs. Elle permet de fournir une borne supérieure pour la matrice de confusion du vote de majorité grâce à la proposition 5.1 qui suit.

Mais définissons, tout d'abord, le risque de Gibbs conditionnel $\mathbf{R}_P(G_\rho, p, q)$ et le risque du vote de majorité conditionnel $\mathbf{R}_P(B_\rho, p, q)$:

$$\mathbf{R}_P(G_\rho, p, q) = \mathop{\mathbf{E}}_{\mathbf{x}\sim P_{|y=p}}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=q), \qquad (5.16)$$

$$\mathbf{R}_P(B_\rho, p, q) = \mathop{\mathbf{E}}_{\mathbf{x}\sim P_{|y=p}}\mathbf{I}\left[\operatorname*{argmax}_{c\in Y}\left\{\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right\}=q\right], \qquad (5.17)$$

où $P_{|y=p}$ est la probabilité conditionnelle d'un exemple $\mathbf{x}$ sachant la classe y. Si $p \neq q$, l'équation (5.16) correspond à l'élément d'indice (p, q) dans la matrice de confusion du classifieur de Gibbs $\mathbf{C}_P^{G_\rho}$ et l'équation (5.17) à l'élément (p, q) de la matrice de confusion du vote de majorité $\mathbf{C}_P^{B_\rho}$. Nous pouvons maintenant prouver les deux résultats suivants.

Proposition 5.1 *Soit P un domaine sur $X \times Y$ tel que $Y = \{1, \ldots, Q\}$. Soit $\mathcal{H}$ un ensemble de classifieurs multiclasse de X vers Y. Pour toute distribution ρ sur $\mathcal{H}$, la valeur réelle du risque conditionnel du vote de majorité ρ-pondéré et celui du classifieur de Gibbs sont liés par l'inégalité suivant :*

$$\forall (q, p) \in Y^2,\ \mathbf{R}_P(B_\rho, p, q) \leq Q\mathbf{R}_P(G_\rho, p, q). \tag{5.18}$$

Démonstration. Voir annexe D.1.1. □

Cette proposition implique le résultat suivant portant directement sur la norme opérateur.

Corollaire 5.3 *Soit P un domaine sur $X \times Y$ tel que $Y = \{1, \ldots, Q\}$. Soit $\mathcal{H}$ un ensemble de classifieurs multiclasse de X vers Y. Pour toute distribution ρ sur $\mathcal{H}$, la matrice de confusion réelle du vote de majorité ρ-pondéré $\mathbf{C}_P^{B_\rho}$ et celle du classifieur de Gibbs $\mathbf{C}_P^{G_\rho}$ sont reliées par l'inégalité suivante :*

$$\left\| \mathbf{C}_P^{B_\rho} \right\|_{op} \leq Q \left\| \mathbf{C}_P^{G_\rho} \right\|_{op}. \tag{5.19}$$

Démonstration. Voir annexe D.1.2 □

Ces deux relations sont à mettre en parallèle avec le résultat obtenu dans le cas binaire (proposition 3.1 du chapitre 3). En fait, selon ces relations, une borne en généralisation pour le classifieur de Gibbs implique une majoration du risque du vote de majorité ρ-pondéré à un facteur Q près. Lorsque le nombre de classe Q tend vers l'infini, cette relation rend le résultat imprécis en particulier lorsque les erreurs des votants individuels sont en moyenne supérieures à $\frac{1}{Q}$. Néanmoins, nous allons voir par la suite qu'il est possible de généraliser la C-borne du théorème 3.3 du chapitre 3 au cadre de la classification multiclasse.

5.3.2 La C-borne en classification multiclasse

La C-borne, en mettant en jeu les premier et second moments de la marge, permet d'avoir une majoration de l'erreur du vote de majorité bien plus précise et plus informative. C'est pourquoi, nous proposons dans cette section de la généraliser au cadre multiclasse en revenant à la mesure d'erreur usuelle avec la fonction de perte $0-1$.

Notions de "marge" en multiclasse

Étant donnée une distribution ρ sur un ensemble $\mathcal{H}$ de votants multiclasse, nous rappelons que l'erreur du vote de majorité ρ-pondéré $\mathbf{R}_P(B_\rho)$ est définie comme la probabilité

qu'il commette une erreur sur un exemple tiré selon le domaine P :

$$\mathbf{R}_P(B_\rho) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathbf{I}\Big(B_\rho(\mathbf{x}) \neq y\Big).$$

Une notion importante liée au vote de majorité est la notion de marge réalisée sur un exemple $(\mathbf{x}, y)$. Il existe différentes façons d'exprimer une telle notion en multiclasse. Nous en présentons trois versions toutes équivalentes en classification binaire.

Tout d'abord, nous faisons appel à la notion de marge multiclasse proposée par [Breiman, 2001] pour les forêts aléatoires.

Définition 5.3 (ρ-marge) *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de votants multiclasse. Étant donnée une distribution ρ sur $\mathcal{H}$, la ρ-marge du vote de majorité $B_\rho(\cdot)$ réalisée sur un exemple $(\mathbf{x}, y)$ tiré selon P est :*

$$\mathcal{M}^\rho(\mathbf{x}, y) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right).$$

Similairement au cadre de la classification binaire présenté dans le chapitre 3, le vote de majorité $B_\rho(\cdot)$ classe correctement un exemple si sa ρ-marge est strictement positive :

$$\mathbf{R}_P(B_\rho) = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y) \leq 0\right). \tag{5.20}$$

Notons que lorsque $Y = \{-1, +1\}$, nous retrouvons la définition usuelle de la marge :

$$\begin{aligned}
\mathcal{M}^\rho(\mathbf{x}, y) &= \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \\
&= \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) \neq y) \\
&= \mathop{\mathbf{E}}_{h\sim\rho} \left[\mathbf{I}(h(\mathbf{x}) = y) - \mathbf{I}(h(\mathbf{x}) \neq y)\right] \\
&= \mathop{\mathbf{E}}_{h\sim\rho} yh(\mathbf{x}) \\
&= y \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}).
\end{aligned}$$

Le fait que la définition de la ρ-marge contienne un terme lié à un maximum, la rend parfois difficile à manipuler. Nous allons donc considérer la relaxation de [Breiman, 2001] qui se base sur la notion de force d'une classe pour un exemple $(\mathbf{x}, y)$ donné. Cette notion est reliée à l'écart entre la classification correcte et les mauvaises classifications indépendamment.

Définition 5.4 (ρ-force) *Soit P un domaine sur $X \times Y$, soit $\mathcal{H}$ un ensemble de classifieurs multiclasse de X vers Y. Étant donnée une distribution ρ sur $\mathcal{H}$, la ρ-force du vote de majorité $B_\rho(\cdot)$ réalisée sur un exemple $(\mathbf{x}, y)$ tiré selon P pour une classe $c \in Y$ est :*

$$\mathcal{S}^\rho(c, (\mathbf{x}, y)) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c).$$

D'après cette définition et l'équation (5.20) on a :

$$\begin{aligned}
\mathbf{R}_P(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathcal{M}^\rho(\mathbf{x},y)\leq 0\right)\\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\exists c\in Y, c\neq y : \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right)\\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\bigvee_{c=1,\ c\neq y}^{Q}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right)\\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\bigvee_{c=1}^{Q}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\wedge c\neq y\right]\right)\\
&\leq \sum_{c=1}^{Q}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\wedge c\neq y\right)\\
&\leq \sum_{c=1}^{Q}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right)\\
&= \sum_{c=1}^{Q}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\leq 0\right)\\
&= \sum_{c=1}^{Q}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathcal{S}^\rho(c,(\mathbf{x},y))\leq 0\right). \qquad (5.21)
\end{aligned}$$

Lorsque $Y=\{-1,+1\}$, on peut trivialement prouver l'égalité entre $\mathbf{R}_P(B_\rho)$ et $\sum_{c=1}^{Q}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathcal{S}^\rho(c,(\mathbf{x},y))\leq 0\right)$.

Enfin, pour relâcher la ρ-marge nous pouvons, en outre, considérer la perte suivante.

Définition 5.5 (la ω-perte) *Soit P un domaine sur $X\times Y$, soit $\mathcal{H}$ un ensemble de classifieurs multiclasse de X vers Y et soit $\omega\in[0,1]$ une constante. Pour toute distribution ρ sur $\mathcal{H}$, on appelle la ω-perte associée à ρ sur un exemple $(\mathbf{x},y)$ tiré selon P, la fonction de perte $\ell^\rho(\omega,(\mathbf{x},y))$ définie par :*

$$\ell^\rho(\omega,(\mathbf{x},y))=\mathbf{I}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq\omega\right]. \qquad (5.22)$$

La valeur réelle de la ω-perte de ρ sur P est :

$$\begin{aligned}
\ell^\rho_P(\omega) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\ell^\rho(\omega,(\mathbf{x},y))\\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\mathbf{I}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq\omega\right]. \qquad (5.23)
\end{aligned}$$

Son estimation empirique calculée sur un échantillon S tiré i.i.d. selon P est :

$$\begin{aligned}
\ell^\rho_S(\omega) &= \frac{1}{m}\sum_{(\mathbf{x},y)\in S}\ell^\rho(\omega,(\mathbf{x},y))\\
&= \frac{1}{m}\sum_{(\mathbf{x},y)\in S}\mathbf{I}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq\omega\right]. \qquad (5.24)
\end{aligned}$$

Cette perte peut être vue comme une relaxation linéaire de la ρ-marge et pour toute distribution ρ sur $\mathcal{H}$, on peut relier l'erreur de $B_\rho(\cdot)$ et la ω-perte associée à ρ grâce au théorème suivant.

Théorème 5.3 *Soit $Q \geq 2$ le nombre de classes. Pour tout domaine P sur $X \times Y$, pour tout exemple $(\mathbf{x}, y)$ tiré i.i.d. selon P et pour toute distribution ρ sur un ensemble de votants multiclasse $\mathcal{H}$, on a :*

$$\ell_P^\rho(\tfrac{1}{Q}) \leq \mathbf{R}_P(B_\rho) \leq \ell_P^\rho(\tfrac{1}{2}). \tag{5.25}$$

Démonstration. Voir annexe D.4. □

Il existe donc une zone d'indécision lorsque $\omega \in \left[\frac{1}{Q}, \frac{1}{2}\right]$ (voir la figure 5.1), et ω doit être choisi avec précaution.

Notons que lorsque $Y = \{-1, +1\}$, $\mathbf{R}_P(B_\rho)$ et $\ell_P^\rho(\frac{1}{2})$ sont trivialement égaux. Les trois mesures de marges présentées sont donc toutes équivalentes lorsque l'on se place dans le cadre de la classification binaire. Cependant, elles diffèrent par l'information qu'elles prennent en considération.

- La ρ-marge est associée à la vraie zone de décision en classification multiclasse et est indépendante de la vraie classe de l'exemple.
- La ρ-force dépend de la vraie classe y de l'exemple et correspond à une combinaison des marges binaires (une classe contre une autre classe) pour les classes $y' \neq y$.
- La ω-perte dépend elle aussi de la vraie classe y, mais ne considère pas les autres classes. C'est une mesure linéaire en fonction de y, plus simple à manipuler mais qui implique, en ce sens, une grande zone d'indécision (voir théorème 5.3).

Ces propriétés sont illustrées sur la figure 5.1 et mènent aux trois généralisations de la C-borne que nous énonçons maintenant.

Généralisations multiclasse de la C-borne

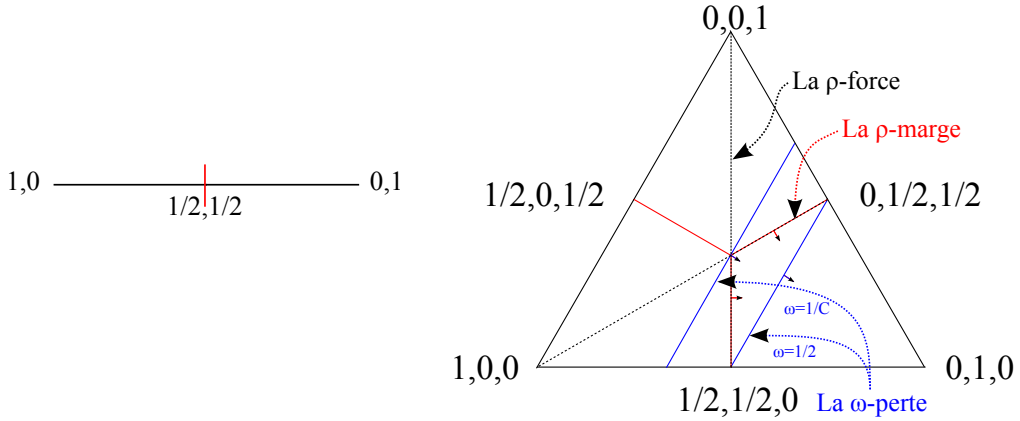
La borne suivante se base sur la définition 5.3 classique de la ρ-marge en multiclasse.

Théorème 5.4 (La C-borne multiclasse) *Pour toute distribution ρ sur une classe de fonctions $\mathcal{H}$ et pour tout domaine P sur $X \times Y$, tel que $\mathcal{M}_P^\rho > 0$, on a :*

$$R(B_\rho) = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathcal{M}^\rho(\mathbf{x}, y) \leq 0\right) \leq C_P^\rho, \tag{5.26}$$

où C_P^ρ est égale à :

$$\begin{aligned} C_P^\rho &= \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x}, y)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\left(\mathcal{M}^\rho(\mathbf{x}, y)\right)^2} \\ &= 1 - \frac{\left(\mathcal{M}_P^\rho\right)^2}{\mathcal{M}_P^{\rho^2}}, \end{aligned}$$



(a) *Les trois mesures de marges sont équivalentes lorsque* $Y = \{-1, +1\}$. *Le codage de la classe* -1 *est* $(1, 0)$, *celui de la classe* $+1$ *est* $(0, 1)$.

(b) *Les trois mesures de marges lorsque* $Y = \{1, 2, 3\}$ *et que la vraie classe de* $\mathbf{x}$ *est* 2. *Le codage de la classe* 1 *est* $(1, 0, 0)$, *celui de la classe* 2 *est* $(0, 1, 0)$, *celui de la classe* 3 *est* $(0, 0, 1)$.

FIGURE 5.1 – *Étant donné un exemple* $(\mathbf{x}, y)$, *on peut représenter le vote* $B_\rho(\mathbf{x})$ *par la combinaison convexe en coordonnées barycentriques où chaque angle correspond à une classe de* $Y = \{1, \ldots, Q\}$. *Les coordonnées de* $B_\rho(\mathbf{x})$ *correspondent alors à* $(\mathbf{E}_{h\sim\rho}\, \mathbf{I}(h(\mathbf{x}) = 1), \ldots, \mathbf{E}_{h\sim\rho}\, \mathbf{I}(h(\mathbf{x}) = Q))$.

où $\mathcal{M}_P^\rho$ *et* $\mathcal{M}_P^{\rho^2}$ *sont respectivement les premier et second moments de la* ρ*-marge* $\mathcal{M}^\rho(\mathbf{x}, y)$ *définis par :*

$$\begin{aligned}
\mathcal{M}_P^\rho &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} \mathcal{M}^\rho(\mathbf{x}, y) \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right),
\end{aligned}$$

$$\begin{aligned}
\mathcal{M}_P^{\rho^2} &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} (\mathcal{M}^\rho(\mathbf{x}, y))^2 \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \right]^2.
\end{aligned}$$

Démonstration. Le résultat se prouve selon le même principe que la C-borne binaire, voir annexe D.2. □

Cette borne offre une relation précise entre l'erreur du vote de majorité et celle du classifieur de Gibbs en ne mettant pas en jeu le nombre de classes Q. Cependant, à des fins algorithmiques, la dérivation d'un algorithme de type MinCq (section 3.4 du chapitre 3) ou P-MinCq (chapitre 4) n'est pas aussi simple qu'en classification binaire. En effet, en pratique, la mise en œuvre des astuces de MinCq est rendue difficile, voir impossible, à cause des termes $\max_{c\in Y, c\neq y} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c)$. Nous énonçons dans la suite les approches liées aux relaxations présentées précédemment.

Ainsi, à l'aide la définition 5.4 de la ρ-force et selon le même principe que la démonstration de la C-borne on obtient la relation suivante :

Théorème 5.5 *Pour toute distribution ρ sur une classe de fonctions $\mathcal{H}$ et pour toute distribution P sur $X \times Y$, tel que $\forall c \in Y, \mathcal{S}_P^\rho(c) > 0$ on a :*

$$R(B_\rho) \leq \sum_{c=1}^{Q} \Pr_{(\mathbf{x},y)\sim P} \left(\mathcal{S}^\rho(c,(\mathbf{x},y)) \leq 0\right) \leq F_P^\rho,$$

où :

$$\begin{aligned} F_P^\rho &= \sum_{c=1}^{Q} \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathcal{S}^\rho(c,(\mathbf{x},y))\right)}{\mathbf{Var}_{(\mathbf{x},y)\sim D'}\left(\mathcal{S}^\rho(\mathbf{x},y,c)\right) - \mathbf{E}_{(\mathbf{x},y)\sim P}\left(\mathcal{S}^\rho(c,(\mathbf{x},y))\right)^2} \\ &= \sum_{c=1}^{Q} \left(1 - \frac{\left(\mathcal{S}_P^\rho(c)\right)^2}{\mathcal{S}_P^{\rho^2}(c)}\right), \end{aligned}$$

où $\mathcal{S}_P^\rho(c)$ et $\mathcal{S}_P^{\rho^2}(c)$ sont respectivement les premier et second moments de la ρ-force de la classe c et sont définis par :

$$\begin{aligned} \mathcal{S}_P^\rho(c) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathcal{S}^\rho(c,(\mathbf{x},y)) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \\ \mathcal{S}_P^{\rho^2}(c) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathcal{S}^\rho(\mathbf{x},y,c)\right)^2 \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right)^2 \end{aligned}$$

Démonstration. Provient de l'inégalité 5.21. □

Cette relation ne contient plus de terme lié au calcul d'un maximum, mais dépend explicitement du nombre de classes Q. Elle peut être vue comme une somme de C-borne pour chaque classe. Ainsi, un inconvénient pratique est qu'il faut résoudre conjointement Q problèmes de minimisation de type MinCq et rend complexe la dérivation d'un algorithme simple.

Finalement, la borne liée à la ω-perte est :

Théorème 5.6 *Pour toute distribution ρ sur la classe de fonctions $\mathcal{H}$ et tout domaine P sur $X \times Y$, si $\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega \right) > 0$, alors :*

$$\ell_P^\rho(\omega) \leq \Omega_P^\rho,$$

où Ω_P^ρ est égal à :

$$\Omega_P^\rho = \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)^2}$$
$$= 1 - \frac{\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)\right]^2}{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)^2}. \tag{5.27}$$

Démonstration. Voir annexe D.3. □

Puisque le terme $\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)$ est linéaire, la dérivation d'un algorithme de type MinCq est plus aisée : on peut facilement fixer le numérateur et minimiser le dénominateur de la borne (5.27). Cependant, les expérimentations préliminaires que nous avons menées (en faisant varier la valeur de ω) ne nous permettent pas encore d'être plus performant qu'un SVM multiclasse. D'une part, le terme linéaire ne capture pas suffisamment bien l'information nécessaire à une classification correcte. D'autre part, les garanties de consistance de MinCq font appel à un ensemble auto-complémenté de votants permettant de s'affranchir du terme de complexité $\mathrm{KL}(\rho\|\pi)$. En classification multiclasse, nous n'avons pas encore réussi à définir la notion d'auto-complémentation.

5.4 Synthèse

Dans ce chapitre, nous avons proposé une analyse PAC-Bayésienne originale pour le problème de la classification multiclasse.

La première contribution repose sur la matrice de confusion utilisée comme une mesure de risque. Mixée avec la norme opérateur sur les matrices, nous avons été capable de prouver, dans le théorème 5.1, une borne en généralisation PAC-Bayésienne sur la "taille" de la matrice de confusion du classifieur de Gibbs. L'idée est que plus la norme est faible, plus le modèle appris sera pertinent pour la tâche considérée. La dérivation de notre résultat tire bénéfice d'une inégalité de concentration proposée par [Tropp, 2011] sur la somme de matrices aléatoires auto-adjointes, que nous avons adaptée à des matrices carrées non auto-adjointes. Un point intéressant est que la borne dépend de la valeur minimale m_- d'exemples d'apprentissage appartenant à une même classe, pour un nombre de classes fixé. Si cette valeur augmente, autrement dit, si le nombre d'exemple d'apprentissage augmente, alors la matrice de confusion du classifieur de Gibbs tend à se rapprocher de sa valeur réelle. Si m_- est petit, alors on voit apparaître le défaut de ce résultat : si une classe est très minoritairement présente, la borne va se dégénérer. De plus, la borne dépend du nombre de classes, cependant, notons que de récents résultats sur les matrices aléatoires proposent des inégalités de

concentration indépendantes de la dimension des matrices (c'est-à-dire du nombre de classes lorsque l'on parle de la matrice de confusion). Nous devrions pouvoir améliorer notre borne en appliquant le résultat de [Hsu *et al.*, 2012].

Alors que la première contribution se focalise sur le classifieur de Gibbs, la seconde explicite différents liens qui existent entre classifieur de Gibbs et vote de majorité ρ-pondéré dans le cadre multiclasse. Nous avons, d'une part, prouvé que la relation triviale repose sur un facteur multiplicateur égal au nombre de classes et, d'autre part, généralisé la C-borne. Cependant, à ce stade, nous n'avons pu dériver d'algorithme à l'image de MinCq qui minimise la C-borne en classification binaire. Nous aimerions donc généraliser la notion d'auto-complémentation de l'ensemble d'hypothèses au multiclasse, mais aussi proposer une mesure de marge/confiance à la fois simple à manipuler et riche en expressivité. Une stratégie envisagée serait de s'orienter vers l'utilisation des codes correcteurs d'erreurs [6] [Peterson et Jr., 1972].

Dans cette partie, nous nous sommes focalisés sur l'apprentissage de votes de majorité pondérés sur un ensemble de votants avec l'approche PAC-Bayésienne dans le cadre classique de la classification supervisée binaire, puis multiclasse. Les contributions de la partie suivante se placent, quant à elles, dans le contexte de l'adaptation de domaine présenté dans le chapitre 2 pour lequel on suppose que les distributions des données de test et d'apprentissage sont différentes. Pour ce faire, nous allons mettre de côté, le temps d'un chapitre, l'analyse PAC-Bayésienne pour étudier le problème de l'adaptation de domaine lorsque l'on veut apprendre un vote de majorité sur un ensemble de similarités (ϵ,γ,τ)-bonnes (présentées en section 1.4.3, en chapitre 1). Enfin, nous retournerons à la théorie PAC-Bayésienne en dérivant la première analyse PAC-Bayésienne pour problème de l'adaptation de domaine.

6. *ECOC : Error-Correcting Output Codes* en anglais.

Troisième partie

Contributions en adaptation de domaine

6 Adaptation de domaine par pondération de fonctions de similarité (ϵ,γ,τ)-bonnes

La partie précédente se positionnait dans le cadre usuel de la classification supervisée où les données d'apprentissage sont représentatives des données à traiter. Cependant, dans de nombreuses applications, cette hypothèse idéale ne peut être vérifiée. Dans de telles situations, une des stratégies vise à adapter un classifieur d'un domaine source vers un domaine cible différent, comme présenté dans le chapitre 2. Dans ce chapitre, nous énonçons notre première contribution en adaptation de domaine. Elle s'inscrit dans la catégorie des methodes de recherche d'un espace de représentation commun et pertinent pour les deux domaines. L'idée est d'utiliser le cadre offert par les fonctions de similarité (ϵ,γ,τ)-bonnes proposé par

[Balcan *et al.*, 2008a, Balcan *et al.*, 2008b] et présenté en section 1.4.3 du chapitre 1. Plus précisément, nous travaillons sur le ϕ^R-espace de projection, défini par les similarités $K(\cdot, \mathbf{x}'_j)$ aux points raisonnables d'un ensemble $R = \{(\mathbf{x}'_j)\}_{j=1}^r$. Pour rappel cet espace est défini par :

$$\phi^R : \left\{ \begin{array}{ccc} X & \to & \mathbb{R}^r \\ \mathbf{x} & \mapsto & (K(\mathbf{x}, \mathbf{x}'_1), \ldots, K(\mathbf{x}, \mathbf{x}'_r))^\top . \end{array} \right.$$

En utilisant la flexibilité de ce cadre, notre méthode construit un espace où les domaines source et cible sont proches tout en gardant de bonnes garanties en généralisation sur le domaine source. Ceci est réalisé à l'aide d'une repondération des similarités, contrôlée par un terme de régularisation, pour rapprocher les exemples sources des exemples cibles. Notre première contribution, décrite en section 6.1, s'inscrit dans le contexte de l'adaptation de domaine non supervisée, c'est-à-dire sans étiquette cible, et requiert des couples d'exemples source-cible non étiquetés à rapprocher. Lorsque ces couples ne sont pas connus, cette approche souffre de la difficulté de leur choix et requiert une étape de paramétrisation basée sur le principe de la validation inverse présentée en section 2.3.3 du chapitre 2. Cette méthode étant très coûteuse, nous proposons une approche itérative, en section 6.2, afin de diminuer le coût de la recherche des couples. Dans un second temps, en section 6.3, nous étendons cette approche au cadre de l'adaptation de domaine semi-supervisée en autorisant l'utilisation de quelques étiquettes cibles. Enfin nous expérimentons les deux approches en section 6.4. Au delà de ces aspects purement techniques, l'intérêt de cette contribution est de proposer une approche permettant de rapprocher les distributions marginales tout en optimisant l'erreur sur le domaine source ce qui, en général, se réalise en deux étapes distinctes.

Les travaux de ce chapitre ont été publiés dans le journal KAIS [Morvant *et al.*, 2012b], dans les conférences ICDM 2011 [Morvant *et al.*, 2011c], CAp 2011 [Morvant *et al.*, 2011a] et CAp 2012 [Morvant *et al.*, 2012a], ainsi que dans le workhop SIMBAD[1] 2011 [Morvant *et al.*, 2011b]. Il a de plus donné lieu à une communication scientifique au workshop de NIPS 2011 *Domain Adaptation Workshop*[2].

6.1 DASF : un algorithme d'adaptation de domaine non supervisée

Pour commencer, nous rappelons quelques notations.

6.1.1 Rappel du cadre de l'adapation de domaine non supervisée

Soit $X \in \mathbb{R}^d$ l'espace de description des données et $Y = \{-1; +1\}$ l'ensemble d'étiquettes. En adaptation de domaine non supervisée, on note P_S le domaine source

1. *Similarity-Based Pattern Analysis and Recognition*, `http://www.dais.unive.it/~simbad/2011/`.
2. `https://sites.google.com/site/nips2011domainadap/home`.

et P_T le domaine cible sur $X \times Y$ (D_S et D_T étant les distributions marginales sur X respectives). On pose $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m^s}$ l'échantillon d'apprentissage étiqueté source constitué de m^s exemples tirés *i.i.d.* selon P_S, $S_u = \{\mathbf{x}_i^s\}_{i=1}^{m_u^s}$ l'échantillon d'apprentissage non étiqueté source constitué de m_u^s exemples tirés *i.i.d.* selon D_S et $T_u = \{\mathbf{x}_i^t\}_{i=1}^{m_u^t}$ l'échantillon d'apprentissage non étiqueté cible constitué de m_u^t exemples tirés *i.i.d.* selon D_T. Étant donnée une classe d'hypothèses $\mathcal{H}$ de X vers Y, l'objectif est de trouver l'hypothèse minimisant l'erreur sur le domaine cible $\mathbf{R}_{P_T}(\cdot)$, sachant que les seules informations d'étiquetage accessibles sont issues du domaine source.

Pour répondre à cette tâche, les bornes d'adaptation de domaine énoncées dans les théorèmes[3] 2.3 et 2.5 du chapitre 2 indique que l'erreur réelle sur le domaine cible est bornée par la somme de trois termes : (A) l'erreur réelle sur le domaine source, (B) la divergence entre les distributions marginales et (C) un terme explicitement lié aux étiquettes sources et cibles. Nous rappelons que l'intuition portée par ces bornes est la suivante : en supposant que (C) soit faible ou négligeable, c'est-à-dire qu'il existe un lien étroit entre les deux domaines, nous voulons trouver un espace commun dans lequel la $\mathcal{H}$-divergence entre les distributions marginales est faible (B), c'est-à-dire qu'elles tendent à être indiscernables, tout en montrant de bonnes performances sur le domaine source (A). Nous nous focalisons sur la borne (2.2) du théorème 2.3 et suivons cette intuition en exploitant les spécificités et la flexibilité de l'espace de projection explicite de la théorie de [Balcan *et al.*, 2008a, Balcan *et al.*, 2008b] pour combiner des fonctions de similarité (ϵ,γ,τ)-bonnes (section 1.4.3, chapitre 1). Concrètement, nous co-régularisons l'apprentissage du classifieur-SF à la lumière de l'information portée par l'échantillon cible T_u afin de minimiser (A) et (B). Nous rappelons que résoudre le problème[4] (1.14) d'apprentissage d'un classifieur-SF à partir de fonctions de similarité (ϵ,γ,τ)-bonnes sur le domaine source permet d'apprendre un classifieur linéaire performant sur le domaine source. Ce classifieur est appris dans le ϕ^R-espace de projection défini explicitement par les similarités à l'ensemble des points raisonnables R. Ce processus implique alors une minimisation naturelle de (A). Pour minimiser la $\mathcal{H}$-divergence (B) entre les marginales D_S et D_T, nous cherchons une fonction de projection $\phi^R(\cdot)$ pour rapprocher D_S et D_T tout en gardant de bonnes performances sur le domaine source, c'est-à-dire une valeur de (A) raisonnable. C'est ici qu'apparaît la co-régularisation : elle va aider à la sélection de points raisonnables pertinents pour notre objectif. Nous repondérons alors la fonction de similarité, en se basant sur le classifieur

3. Nous rappelons la forme des bornes d'adaptation pour la classification binaire, pour tout $h \in \mathcal{H}$:

$$\mathbf{R}_{P_T}(h) \leq \mathbf{R}_{P_S}(h) + \tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu. \qquad \text{(eq. (2.2), théorème 2.3)}$$

$$\mathbf{R}_{P_T}(h) \leq \mathbf{R}_{D_S}(h_S^*, h) + d_{\mathcal{H}}(D_S, D_T) + \upsilon. \qquad \text{(eq. (2.3), théorème 2.5)}$$

4. Nous rappelons la formulation du problème (1.14) énoncé en section 1.4.3, chapitre 1 :

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} \dfrac{1}{m} \sum\limits_{i=1}^{m} \ell_{hinge}(h(\mathbf{x}_i), y_i)) + \lambda \|\boldsymbol{\alpha}\|_1, \\ \text{avec } \ell_{hinge}(h(\mathbf{x}_i), y_i)) = \left[1 - y_i \sum\limits_{j=1}^{r} \alpha_j K(\mathbf{x}_i, \mathbf{x}'_j)\right]_+ . \end{cases} \tag{1.14}$$

linéaire $\boldsymbol{\alpha} \in \mathbb{R}^{|R|}$ appris, pour induire un nouvel espace. Dans cette section, puisque nous supposons l'information sur les étiquettes cibles non disponible, le dernier terme (C) reste difficile à diminuer. Cependant, nous allons proposer une méthode heuristique tirant partie d'un processus de validation inverse pour essayer de le contrôler.

6.1.2 Le problème d'optimisation

Résoudre le problème (1.14) pour apprendre un classifieur-SF revient à minimiser empiriquement l'erreur source et permet de définir un espace de projection approprié pour le domaine source. En effet, les points raisonnables non pertinents ne seront pas considérés : un poids nul leur est affecté dans la solution $\boldsymbol{\alpha}$. À partir de la notion de $\mathcal{H}$-divergence de la définition 2.4 du chapitre 2, nous proposons un terme de régularisation additionel pour forcer le modèle à renvoyer des sorties similaires pour des couples de points source-cible. Ceci tend à diminuer la $\mathcal{H}$-divergence entre les distributions marginales D_S et D_T. Cette idée n'est pas sans rappeler l'intuition portée par la notion d'algorithme robuste [Xu et Mannor, 2010, Xu et Mannor, 2012] présenté en section 1.3.4, chapitre 1 : *"if a testing sample is similar to a training sample then the testing error is close to the training error"*. Nous rappelons qu'en apprentissage supervisé, pour dériver des garanties en généralisation selon la robustesse, il suffit de s'assurer que pour un point de test proche d'un point d'apprentissage de même étiquette, alors la différence entre les fonctions de perte associées à chacun des points est faible. Intuitivement, nous considérons les points sources comme les points d'apprentissage et les points cibles comme les points de test. Ainsi, l'idée est de co-régulariser l'apprentissage afin de rapprocher des couples de points source-cible et de rendre indiscernables les échantillons source S_u et cible T_u.

En considérant la fonction de perte hinge utilisée par le problème (1.14), pour tout modèle h appris et tout couple $(\mathbf{x}^s, \mathbf{x}^t)$ de points source et cible à rapprocher et supposés de même classe y, nous construisons le terme de co-régularisation de sorte qu'il majore l'écart entre les pertes :

$$\left|\ell_{hinge}(h,(\mathbf{x}^s,y))) - \ell_{hinge}\big(h,(\mathbf{x}^t,y)\big)\right| = \left|\left[1 - y\sum_{j=1}^{r}\alpha_j K(\mathbf{x}^s,\mathbf{x}'_j)\right]_+ - \left[1 - y\sum_{j=1}^{r}\alpha_j K(\mathbf{x}^t,\mathbf{x}'_j)\right]_+\right|.$$

Puisque la perte hinge est 1-lipschitzienne ($|[X]_+ - [Y]_+| \leq |X - Y|$), alors :

$$\begin{aligned}\left|\ell_{hinge}(h,(\mathbf{x}^s,y))) - \ell_{hinge}\big(h,(\mathbf{x}^t,y)\big)\right| &\leq \left|\sum_{j=1}^{r}\alpha_j\left(K(\mathbf{x}^s,\mathbf{x}'_j) - K(\mathbf{x}^t,\mathbf{x}'_j)\right)\right| \\ &\leq \sum_{j=1}^{r}\left|\alpha_j\left(K(\mathbf{x}^s,\mathbf{x}'_j) - K(\mathbf{x}^t,\mathbf{x}'_j)\right)\right| \\ &= \left\|(\phi^R(\mathbf{x}^s)^\top - \phi^R(\mathbf{x}^t)^\top)\,\mathrm{diag}(\boldsymbol{\alpha})\right\|_1. \qquad (6.1)\end{aligned}$$

où $\mathrm{diag}(\boldsymbol{\alpha})$ est la matrice diagonale de diagonale $\boldsymbol{\alpha}$. La minimisation du terme de la ligne (6.1) va amener à la sélection de points raisonnables permettant de rapprocher $\mathbf{x}^s$ et $\mathbf{x}^t$ dans le ϕ^R-espace de projection, par conséquent tendant à diminuer la divergence entre les distributions marginales.

À ce stade, nous supposons que les couples de points $(\mathbf{x}^s, \mathbf{x}^t)$ à rapprocher sont connus. Nous expliquerons en section 6.2.1 comment les choisir lorsqu'ils sont inconnus. On pose $\mathcal{C}_{ST} \subseteq S_u \times T_u$ l'ensemble de ces couples de points non étiquetés. Le nouveau terme de régularisation, pondéré par un paramètre β est alors intégré au problème (1.14) pour tous les couples de $\mathcal{C}_{ST}$. Soit $R = \{\mathbf{x}'_j\}_{j=1}^r$ un ensemble de r *landmarks* et $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m^s}$ l'échantillon d'apprentissage de m^s exemples sources étiquetés, notre problème d'optimisation d'adaptation de domaine non supervisé correspond au programme linéaire suivant :

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}) = \dfrac{1}{m^s} \sum\limits_{i=1}^{m^s} \ell_{hinge}(h(\mathbf{x}_i^s), y_i^s) + \lambda \|\boldsymbol{\alpha}\|_1 + \beta \sum\limits_{(\mathbf{x}^s, \mathbf{x}^t) \in \mathcal{C}_{ST}} \left\| \left(\boldsymbol{\phi}^R(\mathbf{x}^s)^\top - \boldsymbol{\phi}^R(\mathbf{x}^t)^\top \right) \operatorname{diag}(\boldsymbol{\alpha}) \right\|_1, \\ \text{avec } \ell_{hinge}(h(\mathbf{x}_i^s), y_i^s) = \left[1 - y_i^s \sum\limits_{j=1}^{r} \alpha_j K(\mathbf{x}_i^s, \mathbf{x}'_j) \right]_+ . \end{cases} \tag{6.2}$$

Ce programme linéaire convexe peut être résolu à l'aide de m^s variables de relachements[5] pour exprimer la perte hinge. Dans ce cas là, le programme aura $O(m^s + r)$ variables et $O(m^s \times r)$ contraintes.

6.1.3 Étude théorique de l'algorithme

Nous proposons dans cette section une étude de la parcimonie et de la capacité en généralisation du problème (6.2).

Pour espérer une bonne adaptation de domaine, comme notre terme de co-régularisation est défini par l'ensemble de couples $\mathcal{C}_{ST}$ et qu'il contribue à trouver un espace de projection approprié, $\mathcal{C}_{ST}$ doit porter de l'information. C'est pourquoi, nous devons supposer la propriété suivante sur les coordonnées des points de $\mathcal{C}_{ST}$ dans le $\boldsymbol{\phi}^R$-espace :

$$\forall \mathbf{x}'_j \in R, \max_{(\mathbf{x}^s, \mathbf{x}^t) \in \mathcal{C}_{ST}} \left| K(\mathbf{x}^s, \mathbf{x}'_j) - K(\mathbf{x}^t, \mathbf{x}'_j) \right| > 0. \tag{6.3}$$

En d'autres termes, pour chaque coordonnées $\mathbf{x}'_j$ dans le $\boldsymbol{\phi}^R$-espace, il existe au moins une coordonnée informative, c'est-à-dire un couple de points dont les similarités $K(\cdot, \mathbf{x}'_j)$ diffèrent. En adaptation de domaine, les domaines étant *a priori* différents, cette hypothèse est tout à fait raisonnable.

Analyse de la parcimonie

Le terme de régularisation $\|\boldsymbol{\alpha}\|_1$ de norme 1 issu du problème (1.14) implique une parcimonie naturelle [Mairal, 2010] du classifieur-SF appris. Notre problème (6.2) comporte un terme supplémentaire à prendre en compte pour l'analyse. Le lemme suivant permet de considérer l'influence de ces deux termes.

5. *Slack variable* en anglais.

Lemme 6.1 *Pour tous les hyperparamètres $\lambda > 0$ et $\beta > 0$, et pour tout ensemble de couples $\mathcal{C}_{ST}$, on pose :*

$$B_R = \min_{\mathbf{x}'_j \in R} \left\{ \max_{(\mathbf{x}^s, \mathbf{x}^t) \in \mathcal{C}_{ST}} \left| K(\mathbf{x}^s, \mathbf{x}'_j) - K(\mathbf{x}^t, \mathbf{x}'_j) \right| \right\}.$$

Si $\boldsymbol{\alpha}^*$ est la solution optimale du problème (6.2)*, alors on a :*

$$\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\beta B_R + \lambda}.$$

Démonstration. Voir annexe E.1. □

Ce lemme montre que la parcimonie du classifieur appris dépend des hyperparamètres λ, β et de la quantité B_R, quantité liée à la distance entre les points des couples de $\mathcal{C}_{ST}$ dans le ϕ^R-espace (B_R est le minimum des déviations maximales des coordonnées des points d'une même couple). Nous pouvons l'interpréter de la manière suivante : plus les distributions marginales D_S et D_T sont éloignées, plus la tâche est dure et plus B_R tend à croître, impliquant alors une plus forte parcimonie. En effet, plus le classifieur est parcimonieux, plus l'espace de projection défini est petit, c'est-à-dire avec moins de coordonnées $K(\cdot, \mathbf{x}')$: on tend à rapprocher les instances sources et cibles plus facilement et avec moins de contraintes. Notons que cette caractéristique peut sembler en contradiction avec l'intuition qui consiste à essayer de trouver un espace le plus expressif possible permettant d'effectuer la tâche d'adaptation. En fait, l'idée est de considérer un ϕ^R-espace de projection suffisamment informatif défini à partir de points raisonnables diversifiés, puis de rapprocher les distributions en réduisant cet espace.

Bornes en généralisation

Nous rappelons que dans le contexte usuel de l'apprentissage supervisé, la robustesse d'algorithme [Xu et Mannor, 2010, Xu et Mannor, 2012] (section 1.3.4, chapitre 1) révèle deux avantages : la prise en considération des termes de régularisation dans la borne et l'étude de contextes non standard tels que l'adaptation de domaine. En gardant à l'esprit que la robustesse d'algorithmes dans le cadre de l'adaptation de domaine est pertinente, nous proposons de dériver une borne de robustesse sur le risque cible de notre algorithme en considérant la $\mathcal{H}$-divergence de la borne d'adaptation de domaine du théorème 2.3 plus adaptée à notre cadre. Tout d'abord, nous prouvons que notre problème d'optimisation (6.2) est robuste sur le domaine source, puis nous en déduisons une borne en généralisation pour le domaine cible.

Théorème 6.1 *Soit (X, ϱ) un espace métrique compact, $K(\cdot,\cdot)$ est une fonction de similarité (ϵ,γ,τ)-bonne continue en son premier argument et $\mathcal{H}$ la classe des classifieurs-SF. Soit les hyperparamètres $\beta > 0$, $\lambda > 0$, l'ensemble des landmarks R et un ensemble de couples $\mathcal{C}_{ST}$ tel que $B_R > 0$. Si l'échantillon d'apprentissage source S est un ensemble de m^s exemples i.i.d. selon le domaine source P_S, alors notre problème* (6.2) *est $\left(2M_\eta, \frac{N_\eta}{\beta B_R + \lambda}\right)$ robuste sur le domaine source P_S, où $\eta > 0$, M_η est le nombre de η-couvertures de X et* $N_\eta = \max_{\substack{(\mathbf{x}^s_a, \mathbf{x}^s_b) \sim (D_S)^2 \\ \varrho(\mathbf{x}^s_a, \mathbf{x}^s_b) \leq \eta}} \left\| \phi^R(\mathbf{x}^s_a)^\top - \phi^R(\mathbf{x}^s_b)^\top \right\|_\infty$.

Démonstration. Voir annexe E.2. □

Dans notre cas, la fonction de perte hinge $\ell_{hinge}(\cdot,\cdot)$ est majorée par une constante notée ℓ^{UP}. Par souci de lisibilité, nous supposerons que $\ell^{UP} = 1$ (ce qui n'est pas vérifié en général, mais qui peut facilement être obtenu via une étape de normalisation lorsque $K(\cdot,\cdot)$ et les α_j sont bornés). On dérive alors du théorème 1.4 la borne en généralisation suivante sur l'erreur source réelle.

Théorème 6.2 *Avec les mêmes notations que celles du théorème 6.1, si $h \in \mathcal{H}$ est le classifieur appris à partir de S en résolvant le problème* (6.2)*, alors pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, on a :*

$$\begin{aligned}\mathbf{R}_{P_S}(h) &\leq \mathbf{R}_{P_S}^{\ell_{hinge}}(h) \\ &\leq \mathbf{R}_S^{\ell_{hinge}}(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2\ln\frac{1}{\delta}}{m^s}}.\end{aligned}$$

Démonstration. D'après le théorème 6.1, le problème (6.2) est $\left(2M_\eta, \frac{N_\eta}{\beta B_R+\lambda}\right)$ robuste sur le domaine source P_S, le résultat est alors obtenu directement en appliquant le théorème 1.4. □

Ce résultat nous permet de prouver la borne en généralisation pour notre approche d'adaptation de domaine non supervisée

Théorème 6.3 *Si $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m^s}$ est un échantillon d'exemples i.i.d. selon le domaine source P_S, si $\mathcal{H}$ est l'espace des classifieurs-SF, et si $h \in \mathcal{H}$ est le classifieur-SF appris à partir de S en résolvant le problème* (6.2)*, alors pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de l'échantillon aléatoire $S \sim (P)^m$, on a :*

$$\begin{aligned}\mathbf{R}_{P_T}(h) &\leq \mathbf{R}_{P_T}^{\ell_{hinge}}(h) \\ &\leq \mathbf{R}_S^{\ell_{hinge}}(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2\ln\frac{1}{\delta}}{m^s}} + \tfrac{1}{2}d_{\mathcal{H}}(D_S, D_T) + \nu,\end{aligned}$$

où ν est l'erreur jointe optimale sur les domaines, $\frac{1}{2}d_{\mathcal{H}}(D_S, D_T)$ est la $\mathcal{H}$-divergence entre les distributions marginales.

Démonstration. La preuve est directe depuis les théorèmes 2.3 et 6.2 : la borne est obtenue en combinant les résultats de robustesse et d'adaptation de domaine. □

Rappelons que les théorèmes 2.1 et 2.2 du chapitre 2 offrent des garanties en généralisation sur l'estimation de $\frac{1}{2}d_{\mathcal{H}}(D_S, D_T)$ par sa version empirique $\frac{1}{2}d_{\mathcal{H}}(S_u, T_u)$ à l'aide respectivement de la VC-dim. et de la complexité de Rademacher. Dans la borne du théorème précédent $\frac{1}{2}d_{\mathcal{H}}(D_S, D_T)$ et ν mesurent la divergence entre les domaines ainsi que la capacité d'adaptation de $\mathcal{H}$. $\mathbf{R}_S^{\ell_{hinge}}(h)$ correspond au risque empirique mesuré sur S l'échantillon d'apprentissage source. La constante $\frac{N_\eta}{\beta B_R+\lambda}$ dépend clairement des termes de régularisations et de N_η. Ce dernier peut être aussi petit que désiré[6],

6. En choisissant η faible et par continuité de la fonction $K(\cdot,\cdot)$ sur son premier argument.

il impliquera alors une croissance de M_η. La borne possède un taux de convergence classique en $O(1/\sqrt{m})$. Une valeur élevée de β, λ ou B_R, signifie que les domaines sont éloignés. Dans notre approche, les termes $\frac{1}{2}d_{\mathcal{H}}(S_u, T_u)$ et $\mathbf{R}_S^{\ell_{linear}}(h)$ sont minimisés par la résolution du problème (6.2). Nous présentons dans la section suivante, notre méthode de sélection des hyperparamètres permettant de garder ces deux termes faibles puis nous introduisons une heuristique nous permettant de faire décroître une estimation du terme ν.

6.1.4 Classifieur inverse et validation des hyperparamètres

Un point crucial pour tout algorithme est le choix des différents hyperparamètres. Pour notre approche, ces hyperparamètres sont λ, β et $\mathcal{C}_{ST}$. Pour les sélectionner nous nous inspirons de la méthode de validation inverse proposée par [Zhong *et al.*, 2010, Bruzzone et Marconcini, 2010] et présentée en section 2.3.3 du chapitre 2. L'idée est que le classifieur, dit inverse, appris à partir de données cibles auto-étiquetées par le classifieur courant doit montrer de bonnes performances sur les données sources lorsque les domaines sont proches. Nous adaptons cette validation inverse en deux points et nous l'illustrons par la figure 6.1. D'une part, elle doit convenir aux fonctions de similarité (ϵ,γ,τ)-bonnes : la validation inverse est explicitement réalisée dans le ϕ^R-espace de projection. D'autre part, aucune information sur les étiquettes cibles n'est disponible : le classifieur inverse est uniquement appris à partir de l'ensemble cible auto-étiqueté. En effet, si les domaines sont suffisamment proches et reliés, alors un tel classifieur doit être efficace sur la tâche source [Bruzzone et Marconcini, 2010]. En d'autres termes, il doit être possible de passer d'un domaine à un autre dans l'espace de projection. Concrètement, notre classifieur inverse h^r correspond au meilleur classifieur-SF appris avec le problème 1.14 — dans le ϕ^R-espace de projection courant — depuis l'échantillon cible $\widehat{T_u} = \{(\mathbf{x}^t, h(\mathbf{x}^t))\}_{\mathbf{x}^t \in T_u}$ auto-étiqueté par le classifieur h appris en résolvant notre problème (6.2) (notons que si l'on dispose de vraies étiquettes cibles, nous pouvons les prendre en compte).

Plus précisément, étant donnés k sous-ensembles de l'échantillon source étiqueté ($S = \cup_{i=1}^{k} S_i$), un classifieur h est appris depuis les $k-1$ sous-ensembles étiquetés et l'échantillon cible non étiqueté en résolvant le problème (6.2), puis le classifieur inverse associé h^r est évalué sur le dernier $k^{\text{ème}}$ sous-ensemble. Son erreur source empirique correspond à la moyenne des erreurs sur les k sous-ensembles :

$$\mathbf{R}_S(h^r) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{R}_{S_i}(h^r).$$

Nous ajoutons à cette approche un élément supplémentaire pour essayer de contrôler le dernier terme de la borne d'adaptation de domaine du théorème 2.3 du chapitre 2, présent dans notre borne en généralisation du théorème 6.3. Nous rappelons que le terme ν, l'erreur jointe définie par $\nu = \mathbf{R}_{P_S}(h^*) + \mathbf{R}_{P_T}(h^*)$ avec $h^* = \text{argmin}_{h\in\mathcal{H}}\ (\mathbf{R}_{P_S}(h) + \mathbf{R}_{P_T}(h))$, peut être associée à la capacité d'adaptation de notre classe d'hypothèses $\mathcal{H}$ dans le ϕ^R-espace de projection. Contrôler ce terme pour

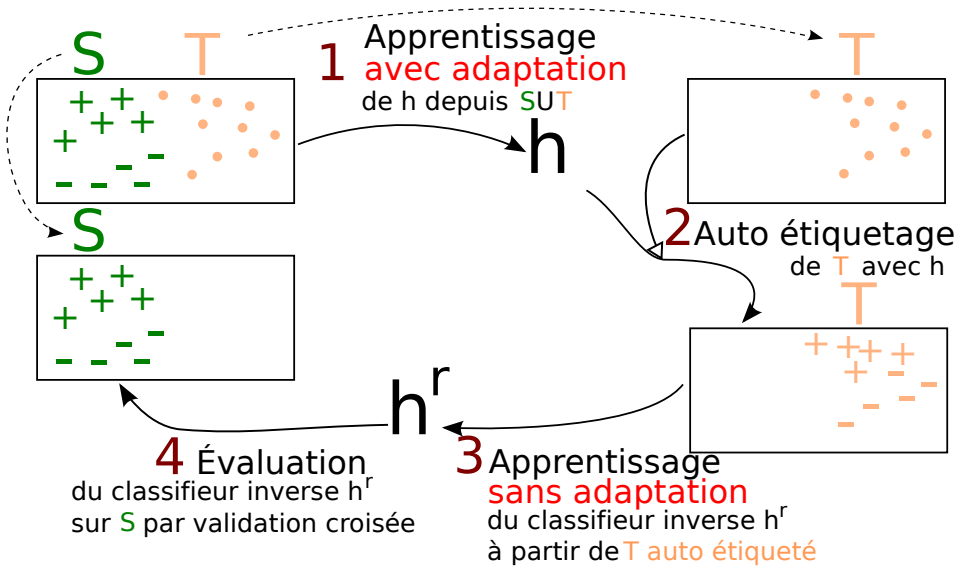


FIGURE 6.1 – *Le processus de validation inverse dans le ϕ^R-espace de projection. Étape* **1** *: Apprendre le classifieur h en résolvant le problème* (6.2)*. Étape* **2** *: Auto-étiqueter l'échantillon cible avec h. Étape* **3** *: Apprendre le classifieur inverse h^r avec le problème* (1.14) *à partir de l'échantillon cible auto-étiqueté. Étape* **4** *: Évaluer h^r sur l'échantillon source.*

sélectionner les hyperparamètres pertinents nous semble judicieux. Cependant, aucune information sur le domaine cible n'est disponible, nous ne pouvons donc pas le calculer. Puisque h^* est clairement relié à la capacité de passer d'un domaine à un autre, nous estimons ν à l'aide du classifieur inverse h^r. Concrètement, à chaque étape de validation inverse, nous divisons en deux l'échantillon cible auto-étiqueté $\widehat{T_u} = \{(\mathbf{x}_i^t, h(\mathbf{x}_i^t))\}_{i=1}^{m_u^t}$: h^r est appris depuis la première puis testé sur la seconde, cette évaluation nous donnant une estimation de l'erreur cible de h^r. En rappelant que l'erreur source de h^r est estimée sur le sous-ensemble courant, l'erreur $\mathbf{R}_S(h^r)$, respectivement $\mathbf{R}_{\widehat{T_u}}(h^r)$, correspond à la moyenne sur les k sous-ensembles de l'estimation de l'erreur source, respectivement de l'erreur cible, de h^r. L'estimation de ν que nous considérons est :

$$\hat{\nu} = \mathbf{R}_S(h^r) + \mathbf{R}_{\widehat{T_u}}(h^r), \tag{6.4}$$

où $\mathbf{R}_{\widehat{T_u}}(h^r)$ est évalué sur l'échantillon cible auto-étiqueté $\widehat{T_u}$. Motivés par la minimisation de la borne d'adaptation de domaine, les hyperparamètres sélectionnés correspondent à ceux associés au $\hat{\nu}$ minimal.

6.2 Simplification de la recherche de l'espace de projection par une pondération itérative

L'un des termes les plus coûteux à estimer par la méthode précédente est l'ensemble des couples $\mathcal{C}_{ST}$. En effet, il dépend généralement de la tâche considérée. Lorsqu'il est inconnu, nous devons le calculer. Néanmoins, aucune étiquette cible n'étant disponible, le couplage de points sources à des points cibles de même classe est donc difficile *a*

priori. Une solution est alors de tester tous les couplages possibles à l'aide de la validation inverse. Les expériences menées dans un tel contexte n'ont cependant produit aucun résultat en un temps raisonnable. Pour contourner cette difficulté, nous proposons une approche itérative basée à la fois sur la sélection d'une quantité limitée de couples et sur une pondération des similarités telle que la distance entre les domaines reste faible. Le critère d'arrêt de notre approche utilise l'estimation de l'erreur jointe optimale $\hat{v}$ précédemment introduit.

6.2.1 Sélectionner les couples $\mathcal{C}_{ST}$

À une itération donnée $\mathfrak{it}$, nous sélectionnons deux sous-ensembles $U_S \subseteq S_u$ et $U_T \subseteq T_u$ de même tailles à l'aide du classifieur inverse $h^r_{\mathfrak{it}-1}$ issu de l'itération précédente et associé au classifieur $h_{\mathfrak{it}-1}$ appris. Ils correspondent aux exemples pour lesquels $h^r_{\mathfrak{it}-1}$ est le plus confiant : pour lesquels la marge est grande. Soit $N \in \mathbb{N}$ et soit δ^H_S, δ^H_T, δ^L_S, δ^L_T un ensemble de paramètres positifs, U_S et U_T tels que $|U_S| = |U_T| \leq N$ sont définis par :

$$U_S = \left\{\mathbf{x}^s \in S_u : \left|h^r_{\mathfrak{it}-1}(\mathbf{x}^s)\right| > \delta^H_S \text{ OU } \left|h^r_{\mathfrak{it}-1}(\mathbf{x}^s)\right| < \delta^L_S\right\},$$
$$\text{et } U_T = \left\{\mathbf{x}^t \in T_u : \left|h^r_{\mathfrak{it}-1}(\mathbf{x}^t)\right| > \delta^H_T \text{ OU } \left|h^r_{\mathfrak{it}-1}(\mathbf{x}^t)\right| < \delta^L_T\right\}.$$

U_S et U_T sont alors utilisés pour construire un couplage biparti $\mathcal{C}_{ST} \subset U_S \times U_T$ minimisant la distance euclidienne dans le nouvel $\phi^R_{\mathfrak{it}}$-espace de projection trouvé à l'itération $\mathfrak{it}$ du processus itératif (décrit dans la section 6.2.2 suivante). Ce problème de couplage biparti particulier peut-être résolu en temps polynomial par le programme quadratique suivant.

$$\begin{cases} \min\limits_{\substack{\chi_{st} \\ 1\leq s\leq |U_S| \\ 1\leq t\leq |U_T|}} \sum\limits_{(\mathbf{x}^s,\mathbf{x}^t)\in U_S\times U_T} \chi_{st} \left\|\phi^R_{\mathfrak{it}}(\mathbf{x}^s) - \phi^R_{\mathfrak{it}}(\mathbf{x}^t)\right\|_2^2 \\ \text{s.c. :} \quad \forall (\mathbf{x}^s,\mathbf{x}^t) \in U_S \times U_T, \quad \chi_{st} \in [0,1], \\ \qquad\qquad \forall \mathbf{x}^s \in U_S, \quad \sum\limits_{\mathbf{x}^t\in U_T} \chi_{st} = 1, \\ \qquad\qquad \forall \mathbf{x}^t \in U_T, \quad \sum\limits_{\mathbf{x}^s\in U_S} \chi_{st} \leq 1. \end{cases} \tag{6.5}$$

$\mathcal{C}_{ST}$ correspond alors aux couples de $U_S \times U_T$ tels que $\chi_{st} = 1$.

Ce problème étant résolu à chaque itération lors de l'étape de validation inverse, nous limitons N pour que le calcul soit rapide et efficace[7]. Dans cette situation, les valeurs de δ^H_S, δ^H_T, δ^L_S, δ^L_T correspondent à celles permettant de sélectionner les N premiers éléments de chaque type.

6.2.2 Un nouvel espace de projection par pondération itérative

Les *landmarks* sélectionnés[8] en résolvant le problème (6.2) définissent un espace de projection dans lequel les domaines tendent à être proches. Nous proposons de ré-

7. Dans les expérimentations menées en section 6.4, nous avons abitrairement posé $N \leq 30$.
8. Les *landmarks* sélectionnés sont ceux associés un α_j non nul.

utiliser les poids α_j pour forcer le nouvel espace de projection à rapprocher les distributions : nous pondérons la fonction de similarité selon $\boldsymbol{\alpha}$. Supposons qu'à l'itération $\mathfrak{it}$, avec la fonction de similarité $K_{\mathfrak{it}}(\cdot,\cdot)$ nous trouvons les nouveaux poids $\boldsymbol{\alpha}^{\mathfrak{it}}$. La fonction $K_{\mathfrak{it}+1}(\cdot,\cdot)$ est alors définie en pondérant $K_{\mathfrak{it}}(\cdot,\cdot)$ vis-à-vis de chacun des *landmarks* $\mathbf{x}'_j \in R$ tel que :

$$\forall \mathbf{x}'_j \in R,\ K_{\mathfrak{it}+1}(\mathbf{x},\mathbf{x}'_j) = \alpha_j^{\mathfrak{it}} K_{\mathfrak{it}}(\mathbf{x},\mathbf{x}'_j).$$

Notons que nous appliquons une normalisation sur $K_{\mathfrak{it}+1}(\cdot,\cdot)$ pour assurer : $K_{\mathfrak{it}+1}(\cdot,\cdot) \in [-1,1]$. Ceci peut donc être vu comme une contraction de l'espace afin de garder une $\mathcal{H}$-divergence empirique faible entre les distributions marginales. En effet, par construction, notre terme de co-régularisation appliqué à l'itération $\mathfrak{it}$ correspond exactement à la minimisation de la distance $\|\cdot\|_1$ dans le nouvel $\phi^R_{\mathfrak{it}+1}$-espace associé à $K_{\mathfrak{it}+1}(\cdot,\cdot)$:

$$\forall (\mathbf{x}^s,\mathbf{x}^t) \in \mathcal{C}_{ST},\ \left\| \phi^R_{\mathfrak{it}+1}(\mathbf{x}^s)^\top - \phi^R_{\mathfrak{it}+1}(\mathbf{x}^t)^\top \right\|_1 = \left\| \left(\phi^R_{\mathfrak{it}}(\mathbf{x}^s)^\top - \phi^R_{\mathfrak{it}}(\mathbf{x}^t)^\top \right) \operatorname{diag}(\boldsymbol{\alpha}^l) \right\|_1 .$$

Ainsi, dans ce nouvel $\phi^R_{\mathfrak{it}+1}$-espace , les points de chacun des couples de $\mathcal{C}_{ST}$ restent naturellement proches. La figure 6.2 illustre cette procédure, que nous itérons à l'étape $\mathfrak{it}+1$ dans le $\phi^R_{\mathfrak{it}+1}$-espace.

Notons que les pondérations possibles dépendent des différents hyperparamètres $\delta^{H/L}_{S/T}$ (liés à $\mathcal{C}_{ST}$) et λ, β du problème (6.2). Une bonne fonction de similarité n'a besoin d'être ni symétrique, ni semi-définie positive. La pondération reste valide si et seulement si la nouvelle similarité est encore suffisamment bonne sur le domaine source. Cette qualité peut être estimée en évaluant ϵ, γ et τ de la définition 1.11 sur $S = \{(\mathbf{x}^s_i, y^s_i)\}_{i=1}^m$ (respectivement notés $\hat{\epsilon}$, $\hat{\gamma}$ et $\hat{\tau}$). En pratique, $\hat{\tau}$ correspond au nombre de *landmarks* sélectionnés par l'algorithme. En conséquence, la $(\hat{\epsilon},\hat{\gamma},\hat{\tau})$-qualité empirique est évaluée par :

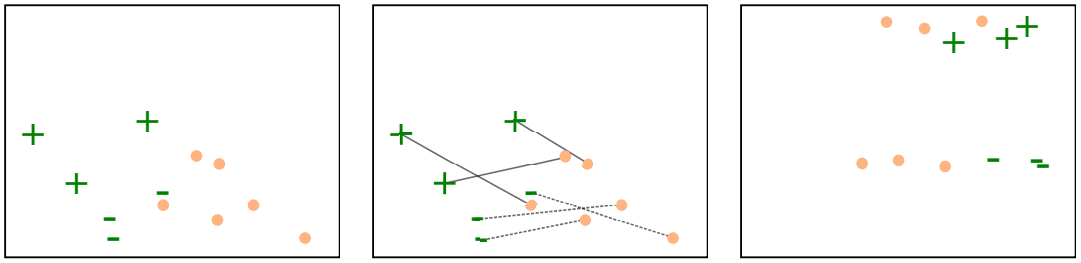
$$\hat{\gamma} = \begin{cases} \gamma_{\max} & \text{si } \underset{\gamma_{\max}>0}{\operatorname{argmax}} \left\{ \forall (\mathbf{x}^s_i, y^s_i) \in S,\ \frac{y^s_i}{r} \sum_{j=1}^r K(\mathbf{x}^s_i, \mathbf{x}'_j) \geq \gamma_{\max} \right\} \text{ existe,} \\ 0 & \text{sinon.} \end{cases}$$

$$\hat{\epsilon} = \begin{cases} 0 & \text{si } \hat{\gamma} > 0, \\ \dfrac{\left| \left\{ (\mathbf{x}^s_i, y^s_i) \in S \ :\ \dfrac{y^s_i}{r} \displaystyle\sum_{j=1}^r K(\mathbf{x}^s_i, \mathbf{x}'_j) < 0 \right\} \right|}{m} & \text{sinon.} \end{cases}$$

Nous nous concentrons donc sur ceux offrant les meilleures $(\hat{\epsilon},\hat{\gamma},\hat{\tau})$-garanties. Concrètement, plus $\hat{\gamma}$ est grand et plus $\hat{\epsilon}$ petit, plus les garanties sont élevées. Notons qu'une une mauvaise similarité impliquera une hausse importante de l'erreur source espérée et ne sera donc pas sélectionnée par la validation inverse.

6.2.3 Critère d'arrêt

Nous rappelons que l'erreur jointe ν est associée à la capacité d'adaptation dans l'espace courant. Le contrôle de sa valeur au cours des itérations apparaît donc comme une solution naturelle à l'arrêt de l'algorithme. D'après la section 6.1.4, pour une itération



(a) *Les points dans le ϕ^R_{it}-espace, $|R| = 2$.*

(b) *Le couplage biparti associé C_{ST} : une ligne représente un couple $(\mathbf{x}^s, \mathbf{x}^t)$.*

(c) *Les points dans le nouvel ϕ^R_{it+1}-espace après la procédure de repondération.*

FIGURE 6.2 – *Une itération it de DASF. Les points sources sont en vert (foncé) (pos. +, neg. −), les points cibles (non étiquetés) sont les ronds orange (clair)*

it donnée, v_{it} est évalué par $\hat{v}_{it} = \mathbf{R}_S(h^r_{it}) + \mathbf{R}_{\widehat{T_u}}(h^r_{it})$, où h^r_{it} est le classifieur inverse associé à h_{it} appris à l'itération it. Si $\hat{v}_{it}$ augmente entre deux itérations, le nouvel espace construit n'est plus approprié et l'espace précédent est préféré.

Le processus s'arrête donc à l'itération it si le $\hat{v}_{it+1}$ suivant a atteint un point de convergence ou a augmenté significativement. Ce critère assure l'arrêt de l'algorithme puisque l'erreur jointe est positive et bornée par 0. L'algorithme 3 décrit l'algorithme de notre méthode d'adaptation de domaine non supervisée (DASF).

Algorithme 3 DASF : *Domain Adaptation with Similarity Function*

entrée fonction de similarité $K(\cdot,\cdot)$, ensemble de *landmarks* R, échantillon source étiqueté S, échantillon source non étiqueté S_u, échantillon cible non étiqueté T_u

sortie classifieur h_{DASF}

$h_0(\cdot) \leftarrow \text{sign}\left[\frac{1}{|R|}\sum_{j=1}^{|R|} K(\cdot, \mathbf{x}'_j)\right]$

$K_1 \leftarrow K$

$it \leftarrow 1$

tant que Le critère d'arrêt n'est pas vérifié **faire**

Sélectionner $U_S \subseteq S_u$, $U_T \subseteq T_u$ avec h^r_{it-1}

$C_{ST} \leftarrow$ Résoudre le problème (6.5)

$\boldsymbol{\alpha}^{it} \leftarrow$ Résoudre le problème (6.2) avec K_{it} et C_{ST}

$K_{it+1} \leftarrow$ Mise à jour de K_{it} en fonction de $\boldsymbol{\alpha}^{it}$

Mise à jour de R (en retirant les landmarks associés à un poids α_j nul)

$it++$

fin tant que

retourner $h_{DASF}(\cdot) = \text{sign}\left[\sum_{x'_j \in R} \alpha^l_j K_{it}(\cdot, \mathbf{x}'_j)\right]$

6.3 SSDASF : extension de DASF à l'adaptation de domaine semi-supervisée

Jusque là, aucune étiquette cible n'était accessible. Dans certaines situations, il est cependant raisonnable de supposer qu'une petite quantité d'étiquettes cibles sont disponibles. Dans ce contexte particulier d'adaptation de domaine semi-supervisée, l'information portée par ces étiquettes cibles peut aider à la recherche du classifieur, ce qui sera confirmé empiriquement dans la section 6.4.

D'après le cadre d'adaptation de domaine semi-supervisée proposé par [Ben-David *et al.*, 2010] et présenté dans la section 2.2.4 du chapitre 2, nous étendons notre approche pour considérer une combinaison linéaire des risques empiriques source et cible. Dans ce cas, l'échantillon d'apprentisssage étiqueté est composé d'un échantillon $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m^s}$ de m^s exemples sources étiquetés *i.i.d.* selon P_S et un échantillon $T = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{m^t}$ de m^t exemples cibles étiquetés *i.i.d.* selon P_T. Soit $\theta \in [0,1]$ tel que $m^t = \theta m$ et $m^s = (1-\theta)m$ impliquant que (S,T) soit constitué de $m = m^t + m^s$ exemples étiquetés tels que $m^t \ll m^s$. Minimiser uniquement l'erreur empirique cible $\mathbf{R}_T(\cdot)$ n'est pas la meilleure solution car T n'est pas représentatif de P_T.

En suivant l'idée de [Ben-David *et al.*, 2010] présentée en section 2.2.4, nous minimisons la combinaison convexe des risques empiriques source et cible présentée en adaptant l'équation (2.5) à la fonction de perte hinge :

$$\kappa\ \mathbf{R}_T^{\ell_{hinge}}(h) + (1-\kappa)\ \mathbf{R}_S^{\ell_{hinge}}(h), \tag{6.6}$$

où $\kappa \in [0,1]$ contrôle le compromis risque cible et risque source. Le risque réel pondéré associé est : $\kappa\ \mathbf{R}_{P_T}^{\ell_{hinge}}(h) + (1-\kappa)\mathbf{R}_{P_S}^{\ell_{hinge}}(h)$.

Nous reformulons alors notre problème d'optimisation précédent (6.2) pour prendre en compte quelques étiquettes cibles. Étant donné (S,T) un échantillon de m exemples, $\mathcal{C}_{ST} \subset S_u \times T_u$ un ensemble de couple et $\kappa \in [0,1]$, nous définissons le problème (6.7) de minimisation suivant. L'algorithme itératif global (SSDASF) est décrit dans l'algorithme 4.

$$\left\{\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}} & (1-\kappa)\dfrac{1}{m^s}\displaystyle\sum_{i=1}^{m^s}\ell_{hinge}(h(\mathbf{x}_i^s), y_i^s) + \kappa\dfrac{1}{m^t}\sum_{i=1}^{m^t}\ell_{hinge}\big(h,(\mathbf{x}_i^t, y_i^t)\big) + \lambda\|\boldsymbol{\alpha}\|_1 \\ & \quad + (1-\kappa)\ \beta \displaystyle\sum_{(\mathbf{x}^s,\mathbf{x}^t)\in\mathcal{C}_{ST}} \left\| \left(\phi^R(\mathbf{x}^s)^\top - \phi^R(\mathbf{x}^t)^\top\right)\mathrm{diag}(\boldsymbol{\alpha})\right\|_1, \\ \text{avec} & \ell_{hinge}(h(\mathbf{x}_i), y_i) = \left[1 - y_i\displaystyle\sum_{j=1}^{r}\alpha_j K(\mathbf{x}_i, \mathbf{x}'_j)\right]_+. \end{array}\right. \tag{6.7}$$

Alors que le problème non supervisé (6.2) se focalise uniquement sur la minimisation du risque empirique source, le problème (6.7) minimise la combinaison convexe des risques empiriques source et cible (équation (6.6)) et peut être vu comme une généralisation du problème (6.2). En effet, un lien existe entre les deux problèmes :

- si $\kappa = 0$, aucune étiquette cible n'est utilisée, nous revenons au problème (6.2) d'adaptation de domaine non supervisée ;

Algorithme 4 SSDASF : *Semi-Supervised Domain Adaptation with Similarity Function*

entrée fonction de similarité $K(\cdot,\cdot)$, ensemble de *landmarks* R, échantillon étiqueté source S, échantillon étiqueté cible T, échantillon source non étiqueté S_u, échantillon cible non étiqueté T_u

sortie classifieur h_{SSDASF}

$h_0(\cdot) \leftarrow \text{sign}\left[\frac{1}{|R|}\sum_{j=1}^{|R|} K(\cdot,\mathbf{x}'_j)\right]$

$K_1 \leftarrow K$

$\mathfrak{it} \leftarrow 1$

tant que le critère d'arrêt n'est pas vérifié **faire**

Sélectionner $U_S \subseteq S_u$, $U_T \subseteq T_u$ avec $h^r_{\mathfrak{it}-1}$

$\mathcal{C}_{ST} \leftarrow$ Résoudre le problème (6.5)

$\boldsymbol{\alpha}^{\mathfrak{it}} \leftarrow$ **Résoudre le problème** (6.7) **avec** $K_{\mathfrak{it}}$ **et** $\mathcal{C}_{ST}$

$K_{\mathfrak{it}+1} \leftarrow$ Mise à jour $K_{\mathfrak{it}}$ selon $\boldsymbol{\alpha}^{\mathfrak{it}}$

Mise à jour de R

$\mathfrak{it}++$

fin tant que

retourner $h_{SSDASF}(\cdot) = \text{sign}\left[\sum_{x'_j \in R} \alpha_j^{\mathfrak{it}} K_{\mathfrak{it}}(\cdot,\mathbf{x}'_j)\right]$

- dans le cas contraire, si $\kappa = 1$, nous tombons dans un cadre d'apprentissage supervisé usuel pour lequel l'échantillon d'apprentissage est tiré selon P_T. Dans ce cas là, le théorème 1.4 de [Xu et Mannor, 2010] peut être prouvée sur P_T.

Nous adaptons l'étude de la parcimonie à notre contexte d'adaptation de domaine semi-supervisée.

Lemme 6.2 *Pour tous les hyperparamètres $\lambda > 0$, $\beta > 0$, $\kappa \in [0,1]$ et pour tout ensemble de couples $\mathcal{C}_{ST}$, on pose :*

$$B_R = \min_{\mathbf{x}'_j \in R} \left\{ \max_{(\mathbf{x}^s,\mathbf{x}^t)\in\mathcal{C}_{ST}} \left| K(\mathbf{x}^s,\mathbf{x}'_j) - K(\mathbf{x}^t,\mathbf{x}'_j) \right| \right\}.$$

Si $\boldsymbol{\alpha}^*$ est la solution optimale du problème (6.7)*, alors on a :*

$$\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{(1-\kappa)\beta B_R + \lambda}.$$

Démonstration. Même processus de preuve que pour le lemme 6.1. □

Ce lemme montre qu'avec des étiquettes cibles additionnelles, c'est-à-dire $(1-\kappa) < 1$, la parcimonie du modèle induit tendra à être moins forte que sans utilisation d'étiquettes cibles.

6.4 Expérimentations

Dans cette section, nous évaluons notre approche DASF, et son extension semi-supervisée SSDASF, sur un problème synthétique et une tâche réelle d'annotation d'images. Premièrement, nous présenterons en section 6.4.1 la fonction de similarité utilisée. Plus précisément, nous proposons une heuristique pour modifier en amont l'espace de projection afin d'obtenir une fonction de similarité (ϵ,γ,τ)-bonne, ni symétrique, ni semi-définie positive, pertinente pour une tâche d'adaptation de domaine. Puis en section 6.4.2, nous introduisons le protocole général de nos expériences. Les résultats sur le jeu de données synthétiques sont présentés en section 6.4.3, ceux pour la tâche de classification d'images en section 6.4.4.

6.4.1 Définir une fonction de similarité (ϵ,γ,τ)-bonne

Nous proposons ici d'introduire un pré-traitement simple pour définir une fonction de similarité non symétrique et non SDP. D'après le résultat théorique de [Ben-David *et al.*, 2010] (théorème 2.3) le classifieur appris doit être performant sur le domaine cible mais aussi sur le domaine source. Ainsi, afin d'aider à l'adaptation sur le domaine cible nous proposons de lier les domaines en considérant l'information portée par l'ensemble des échantillons. Concrètement, nous considérons une fonction de similarité $K_{ST}(\cdot,\cdot)$ construite en normalisant une fonction de similarité donnée $K(\cdot,\cdot)$ sur l'échantillon non étiqueté $ST = (S_u, T_u)$. Ce choix est clairement heuristique et notre but est simplement d'évaluer l'intérêt de la normalisation d'une fonction de similarité pour les problèmes d'adaptation de domaine. Rappelons que la définition 1.11 indique qu'une fonction de similarité doit être (ϵ,γ,τ)-bonne relativement à un ensemble de points raisonnables R. L'idée est donc de la normaliser de sorte que chaque similarité vis-à-vis d'un *landmark* $\mathbf{x}'_j$ de R ait une une moyenne de 0 et une variance unitaire sur ST. Étant donnée une fonction de similarité $K(\cdot,\cdot)$ vérifiant la définition 1.11, notre fonction de similarité normalisée K_{ST} est définie par :

$$\forall \mathbf{x}'_j \in R,\ K_{ST}(.,\mathbf{x}'_j) = \begin{cases} \dfrac{K(.,\mathbf{x}'_j) - \mu_{\mathbf{x}'_j}}{\sigma_{\mathbf{x}'_j}} & \text{si} \quad -1 \leq \dfrac{K(.,\mathbf{x}'_j) - \hat{\mu}_{\mathbf{x}'_j}}{\hat{\sigma}_{\mathbf{x}'_j}} \leq 1\ , \\ -1 & \text{si} \quad \dfrac{K(.,\mathbf{x}'_j) - \hat{\mu}_{\mathbf{x}'_j}}{\hat{\sigma}_{\mathbf{x}'_j}} \leq -1\ , \\ 1 & \text{si} \quad 1 \leq \dfrac{K(.,\mathbf{x}'_j) - \hat{\mu}_{\mathbf{x}'_j}}{\hat{\sigma}_{\mathbf{x}'_j}}, \end{cases} \tag{6.8}$$

où $\hat{\mu}_{\mathbf{x}'_j}$ est la moyenne empirique des similarités à $\mathbf{x}'_j$ sur ST :

$$\forall \mathbf{x}'_j \in R,\ \hat{\mu}_{\mathbf{x}'_j} = \frac{1}{|ST|} \sum_{\mathbf{x} \in ST} K(\mathbf{x},\mathbf{x}'_j),$$

et $\hat{\sigma}_{\mathbf{x}'_j}$ est l'estimateur non biaisé empirique de la variance associée à $\hat{\mu}_{\mathbf{x}'_j}$:

$$\forall \mathbf{x}'_j \in R,\ \hat{\sigma}_{\mathbf{x}'_j} = \sqrt{\frac{1}{|ST|-1} \sum_{\mathbf{x} \in ST} \left(K(\mathbf{x},\mathbf{x}'_j) - \hat{\mu}_{\mathbf{x}'_j}\right)^2}.$$

Par construction la similarité $K_{ST}(\cdot,\cdot)$ n'est alors ni symétrique, ni semi-définie positive.

Dans toutes nos expériences, la fonction de similarité $K(\cdot,\cdot)$ considérée est un noyau gaussien défini dans l'équation (1.12) en chapitre 1. Cependant, en fonction des échantillons, la similarité $K_{ST}(\cdot,\cdot)$ n'offre pas toujours de meilleures (ϵ,γ,τ)-garanties par rapport au noyau gaussien. Par la suite, les résultats reportés correspondent à la fonction produisant les meilleurs résultats. Ceux obtenus avec $K_{ST}(\cdot,\cdot)$ sont indiqués par *. Comme nous le verrons ils correspondent généralement aux tâches d'adaptation de domaine les plus difficiles.

6.4.2 Protocole expérimental

Nous comparons notre algorithme DASF à différentes méthodes : un SVM classique (section 1.4.2), appris uniquement sur le domaine source, l'algorithme transductif de SVM développé pour d'apprentissage semi-supervisé (lorsque des données d'apprentissage ne sont pas étiquetées) [Vapnik, 1998] (TSVM) et la méthode d'adaptation de domaine DASVM [Bruzzone et Marconcini, 2010] (section 2.3.1). Nous considérons un noyau gaussien (1.12) pour ces trois méthodes afin de faciliter la comparaison. Nous avons utilisé la bibliothèque SVM-light [Joachims, 1999] pour SVM et TSVM (les paramètres sont choisis par validation croisée sur les données sources). DASVM est implémenté avec la bibliothèque LibSVM [Chang et Lin, 2001]. Les paramètres de DASVM et DASF sont sélectionnés selon une grille de recherche par validation inverse. De plus, nous nous comparons à un classifieur-SF appris uniquement sur le domaine source. Pour DASF et SF, les *landmarks* sont issus de l'échantillon étiqueté source. D'après le lemme 2.1, nous estimons la $\mathcal{H}$-divergence entre les deux distributions marginales en apprenant un classifieur-SF (qui est un classifieur linéaire dans l'espace de représentation) visant à séparer la tâche source de la tâche cible. Nous notons cette estimation $\frac{1}{2}\hat{d}_{\mathcal{H}}$. Une valeur proche de 0 indique des distributions proches, tandis qu'une valeur proche de 1 indique une tâche d'adaptation difficile. Nous observons aussi l'influence des différents hyperparamètres de notre méthode : λ (β fixé) et β (λ fixé). Les valeurs testées pour ces paramètres sont 0, 0.01, 0.1, 0.25, 0.5, 0.75 et 1.

Dans un second temps, nous étudions le comportement de notre algorithme SSDASF à apprendre un classifieur performant si une partie de l'échantillon étiqueté est issue du domaine cible. Dans ce but, chacune des tâches d'adaptation est répétée 9 fois en utilisant l'extension semi-supervisée SSDASF avec 9 échantillons aléatoires de 2, 4, 8, 10, 12, 14, 16, 18 et 20 exemples cibles étiquetés dans l'échantillon d'apprentissage. De plus, nous étudions l'impact du paramètre de contrôle du compromis entre risque source et risque cible, κ du problème (6.7). Dans ce cas, nous fixons λ, β et la quantité d'étiquettes cibles à 10. Les valeurs de κ testées sont 0, 0.01, 0.1, 0.25, 0.5, 0.75, 0.80, 0.85, 0.90, 0.95, 0.99 et 1. Dans ce cas, nous ajoutons des *landmarks* cibles à R, alors que pour DASF, R ne contient uniquement des exemples sources [9]. Enfin, dans cette étude nous n'avons pas reporté les résultats pour λ et β du problème (6.7), car SSDASF a montré un comportement similaire à DASF.

9. Ce point est discuté en conclusion du chapitre

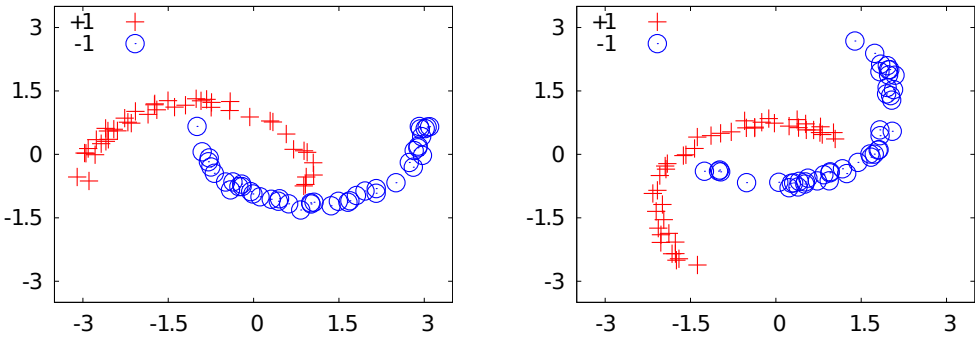


FIGURE 6.3 – **Problème jouet.** *À gauche : un échantillon source. À droite : un échantillon cible avec* 50° *de rotation.*

Parallèlement, nous calculons les coûts en temps moyen de chaque algorithme lorsque les paramètres sont fixés. La durée de l'apprentissage du classifieur-SF est le temps de base. Nous reportons donc le temps d'exécution des autres méthodes comme le rapport en fonction ce temps de base [10]. Nous considérons aussi le temps d'exécution de la première itération de nos approches (noté $_{it_1}$). Nous rappelons que considérer tous les couples possibles est insoluble : l'approche itérative permet de contourner le problème et nous verrons qu'elle reste raisonnablement compétitive en terme de temps d'exécution.

6.4.3 Problème jouet synthétique

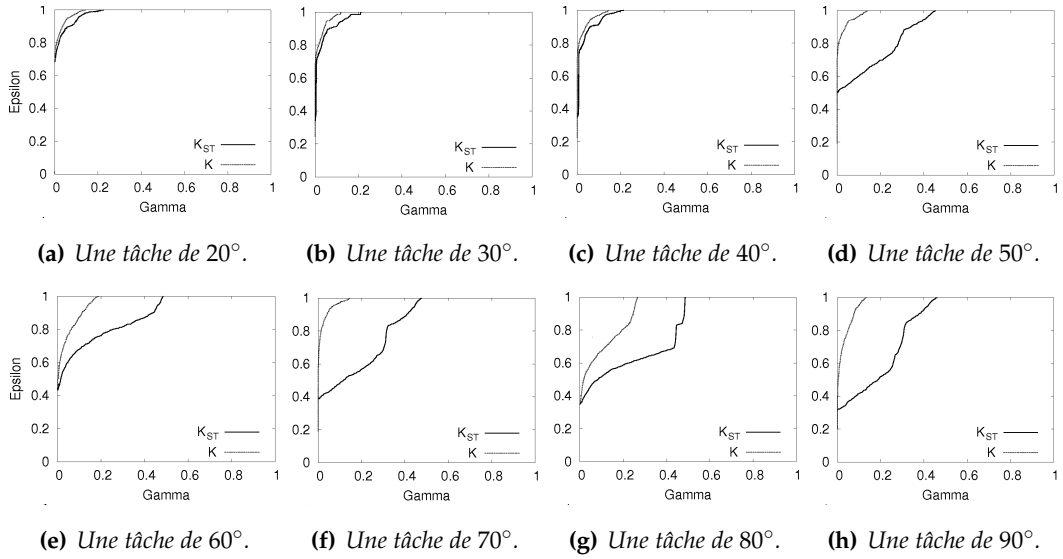
Protocole

Ici, le domaine source correspond à une tâche de classification binaire classique appelée les lunes jumelles (une classe par lune, voir la figure 6.3). Nous considérons 8 domaines cibles différents, chacun produit par une rotation anti-horaire, selon 8 angles, du domaine source. Plus l'angle est grand, plus le problème devient difficile. Pour chaque domaine, nous générons 300 instances (150 de chaque classe). Les algorithmes sont évalués sur un échantillon de test composé de 1 500 exemples tirés selon le domaine cible (et non utilisé par les algorithmes, on dispose aussi d'un échantillon de test source de 1 500 exemples). Chacun des problèmes d'adaptation de domaine est répété 10 fois.

Sélection de la "meilleure" fonction de similarité

Avant de détailler les résultats, nous évaluons si $K_{ST}(\cdot,\cdot)$ est meilleure que $K(\cdot,\cdot)$ sur le domaine cible. D'après la définition 1.11 d'une fonction de similarité (ϵ,γ,τ)-bonne, nous proposons une étude empirique des (ϵ,γ)-garanties sur le domaine cible. Dans ce but, étant donné $R = \{\mathbf{x}'_j\}_{j=1}^{r}$, le paramètre ϵ est estimé, sur un échantillon cible étiqueté $\{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{m^t}$, comme une fonction de γ. En fait, pour chaque valeur de γ, ϵ

10. Par exemple, un coût de 0.5 signifie que l'algorithme a mis la moitié moins de temps que le classifieur-SF et un coût de 2 signifie qu'il lui a fallu deux fois plus de temps.



(a) *Une tâche de* 20°. **(b)** *Une tâche de* 30°. **(c)** *Une tâche de* 40°. **(d)** *Une tâche de* 50°.

(e) *Une tâche de* 60°. **(f)** *Une tâche de* 70°. **(g)** *Une tâche de* 80°. **(h)** *Une tâche de* 90°.

FIGURE 6.4 – **Problème jouet.** *Qualité des fonctions de similarités :* ϵ (Epsilon) *est une fonction de* γ (Gamma).

correspond à la proportion des exemples $\mathbf{x}_t$ verifiant :

$$\sum_{j=1}^{r} \frac{y_j y_i^t K(\mathbf{x}_i^t, \mathbf{x}_j')}{r} < \gamma.$$

Pour chacune des similarités, ϵ est calculé pour 20 valeurs de γ entre 0 et 1, puis représentée par une courbe. Pour chaque angle de rotation, nous obtenons deux courbes et la meilleure fonction de similarité correspond à celle avec la plus faible aire sous la courbe, impliquant une erreur plus faible en moyenne. La qualité des fonctions de similarité pour chacune des tâches est reportée sur la figure 6.4. Pour les tâches les plus difficiles ($\geq$ 50°), $K_{ST}(\cdot,\cdot)$ montre une meilleure qualité que $K(\cdot,\cdot)$. Pour les tâches plus simples, l'amélioration n'est pas significative, justifiant que $K(\cdot,\cdot)$ peut être meilleure. Notre normalisation apparaît pertinente lorsque le problème d'adaptation est difficile.
Notons que le taux de ϵ est relativement élevé s'explique par l'utilisation de *landmarks* issus uniquement du domaine source (permettant d'étudier les capacités d'adaptation). Nous présentons maintenant les résultats obtenus sur les lunes jumelles.

Résultats

Le taux moyen d'erreur de chaque méthode est reporté dans la table 6.1. De plus, nous indiquons le nombre moyen de vecteurs de support (SV) obtenus par SVM, TSVM et DASVM, le nombre de *landmarks* (LAND.) sélectionnés par SF et DASF et une estimation de la $\mathcal{H}$-divergence entre les domaines dans le ϕ_0^R-espace de représentation inital et dans le ϕ_{final}^R-espace final. Nous faisons les remarques suivantes.

- DASF est en moyenne plus performant. Il est significativement meilleur pour tous les problèmes d'angle supérieur à 20°. Dès 60°, la difficulté augmente et

Angle de rotation	20°	30°	40°	50°	60°*	70°*	80°*	90°*
SVM	0.1032	0.2401	0.3116	0.4000	0.5282	0.7388	0.8078	0.8280
SV					18			
SF	0.0760	0.1819	0.2745	0.4215	0.5607	0.6080	0.6407	0.6327
Land.			24		22	20	20	20
TSVM	**0**	0.2102	0.2534	0.2909	0.3528	0.7872	0.8108	0.8251
SV	28	37	37	37	38	35	37	36
DASVM	**0**	0.2159	0.2837	0.3341	0.3843	0.7466	7893	0.8194
SV	20	20	26	28	29	34	38	23
DASF	0.0020	**0.0045**	**0.0897**	**0.1873**	**0.3477**	**0.3805**	**0.3909**	**0.4025**
Land.	**10**	**10**	**9**	**8**	**4**	**4**	**4**	**3**
$\hat{d}_{\mathcal{H}}$ in ϕ_0^R	0.29	0.58	0.65	0.67	0.67	0.65	0.66	0.65
$\hat{d}_{\mathcal{H}}$ in ϕ_{final}^R	0.16	0.33	0.41	0.42	0.19	0.20	0.24	0.22

TABLE 6.1 – **Problème jouet.** *Taux d'erreur pour les* 8 *problèmes.*

les performances de TSVM et DASVM chutent, alors que DASF reste compétitif. Comme nous l'avons vu sur la figure 6.4, nous avons la confirmation que dans une telle situation, la fonction normalisée (*) est préférée.

- Les *landmarks* (LAND.) sont significativement moins nombreux que les vecteurs de support (SV). C'est la confirmation que DASF produit des modèles très parcimonieux avec de bonnes performances. Cette quantité est 3 à 12 fois plus faible. De plus, les classifieurs appris par DASF sont plus parcimonieux que les classifieurs-SF qui font aussi appels à une régularisation de type $\|\cdot\|_1$. Finalement, comme le suggère le lemme 6.1, ils tendent à être plus parcimonieux lorsque le problème est difficile.

- À la dernière itération (entre 1 et 9), la divergence entre les domaines est plus basse. Notre méthode rapproche donc bien les distributions. Cependant, l'algorithme tend à construire un "petit" espace pour les tâches les plus dures. Ceci est probablement dû à la nécessité d'avoir des distributions suffisamment proches, et peut induire une perte de performance et/ou d'expressivité.

La figure 6.5 montre une exécution de DASF sur deux tâches d'adaptation de domaine. Dans les deux cas, la $\mathcal{H}$-divergence empirique décroît significativement en comparaison de la première itération.

L'algorithme s'arrête au moment où l'erreur jointe atteint un minimum, après une décroissance continue. Notons, d'une part, que l'espace de projection final n'est pas nécessairement celui associé à la plus faible divergence et, d'autre part, que l'erreur sur l'échantillon de test source augmente : notre problème d'optimisation cherche le meilleur compromis entre la minimisation de la divergence et celle de l'erreur source, afin d'être le meilleur possible sur le domaine cible. Pour l'exemple de rotation de 30°, DASF construit un classifieur d'erreur nulle sur l'échantillon de test cible. Pour le

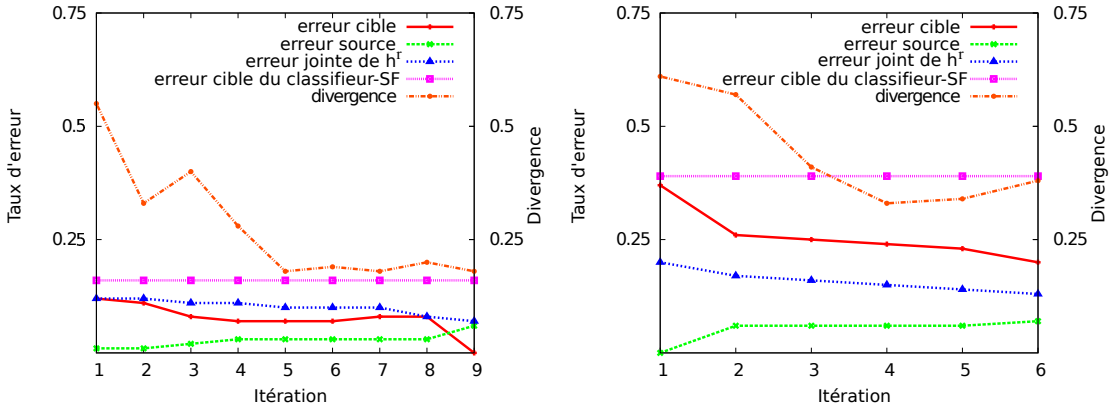


FIGURE 6.5 – **Problème jouet.** *Deux exécutions de DASF. À gauche pour un angle de* 30°*, à droite de* 50°*. Le taux d'erreur est reporté en ordonné à gauche, la mesure de divergence en ordonnée à droite. À chaque itération it (en abscisse), on mesure les taux d'erreurs de h_{it} sur les échantillons de test sources et cibles, la divergence $\hat{d}_{\mathcal{H}}$, l'erreur du classifieur joint, l'erreur sur l'échantillon de test cible d'un classifieur-SF appris sans adaptation.*

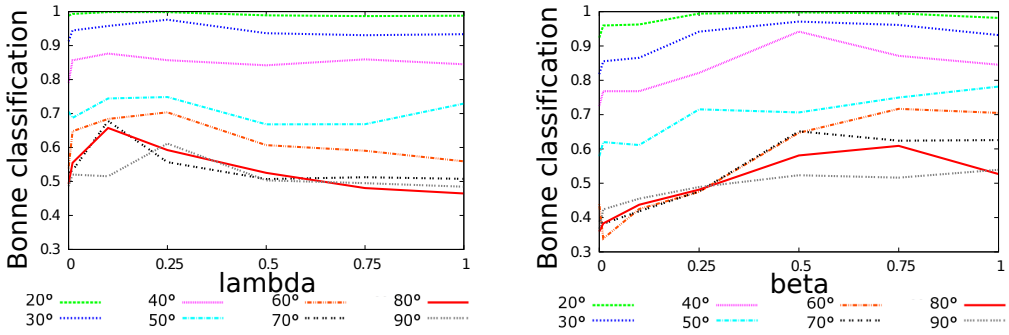
problème de 50°, la difficulté est plus grande et DASF infère un classifieur performant, meilleur que le classifieur-SF appris seulement sur les données sources.

DASF : Influence des hyperparamètres λ et β

Nous observons l'évolution du taux d'erreur en fonction des différents hyperparamètres de notre problème d'optimisation (6.2) pour DASF.
Dans un premier temps, les figures 6.6(a) et 6.6(b) reportent le taux moyen de bonne classification pour chaque angle de rotation en fonction respectivement de de $\lambda \in [0,1]$ (selon le meilleur β) et de $\beta \in [0,1]$ (selon le meilleur λ). Nous observons un gain selon β plus siginificatif que celui de λ. En effet, une bonne valeur de β mène à un gain de 0.05 à 0.35, alors que le gain associé à un λ approprié ne dépasse jamais 0.2. Les valeurs de λ et β menant aux modèles les plus performants appartiennent respectivement à $[0.1, 0.25]$ et $[0.5, 1]$. Précisons que les tâches difficiles sont plus sensibles aux paramètres (pour les plus simples, le gain avec λ est quasiment nul). Dans un second temps, les tables 6.1(a) et 6.1(b) montrent la quantité moyenne de *landmarks* associée aux modèles précédemment considérés. Ici, c'est la combinaison de λ et de β qui infère des modèles plus parcimonieux. De plus comme le suggère le lemme 6.1, la parcimonie augmente avec la valeur de β associée au classifieur le plus performant et avec la difficulté de la tâche (voir la table 6.1(b)).

SSDASF : Influence de données étiquetées cibles

Nous étudions maintenant le comportement de SSDASF, l'extension semi-supervisée de DASF combinant des étiquettes sources et cibles.
Tout d'abord, le taux moyen de bonne classification en fonction de la quantité de données étiquetées cibles m^t est indiquée sur la figure 6.7(a). On y observe le comportement attendu de l'augmentation de la performance avec m^t. Plus la tâche est



(a) *Le taux moyen en fonction de λ avec β=0.75.* **(b)** *Le taux moyen en fonction de β avec λ=0.15.*

FIGURE 6.6 – **Problème jouet.** *Taux de bonne classification en fonction de λ et β avec DASF.*

(a) *La parcimonie en fonction λ avec $\beta = 0.75$.*

Angle	20°	30°	40°	50°	60°	70°	80°	90°
$\lambda = 0$	16	12	13	8	12	6	7	8
$\lambda = 0.01$	15	10	9	6	5	5	8	11
$\lambda = 0.1$	**10**	**9**	**8**	**6**	3	**3**	**2**	6
$\lambda = 0.25$	15	11	12	5	**4**	5	6	**3**
$\lambda = 0.5$	18	17	14	10	6	6	9	4
$\lambda = 0.75$	13	15	13	10	8	9	8	8
$\lambda = 1$	15	21	16	9	12	12	8	7

(b) *La parcimonie en fonction β avec $\lambda = 0.15$.*

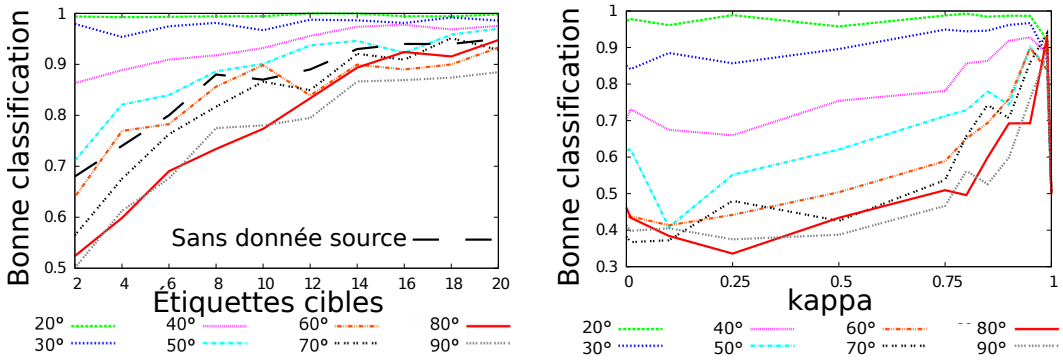
Angle	20°	30°	40°	50°	60°	70°	80°	90°
$\beta = 0$	24	24	24	24	22	20	20	20
$\beta = 0.01$	22	18	20	20	4	6	6	11
$\beta = 0.1$	16	17	17	19	3	4	5	9
$\beta = 0.25$	13	11	13	12	3	3	5	7
$\beta = 0.5$	**11**	**10**	**8**	11	3	**3**	3	7
$\beta = 0.75$	12	16	11	11	**3**	4	**3**	6
$\beta = 1$	14	11	8	**6**	3	5	4	**5**

TABLE 6.2 – **Problème jouet.** *Quantité moyenne de landmarks illustrant la parcimonie. En gras, les valeurs associées au modèle le plus performant.*

dure, plus cette augmentation est significative. Cependant, pour les problèmes les plus difficiles ($\geq 70°$) nous sommes incapables d'inférer un classifieur plus efficace que le classifieur-SF appris uniquement depuis les étiquettes cibles. Ce résultat reste cohérent avec l'intuition selon laquelle les tâches difficiles, pour lesquelles les domaines sont éloignés, requièrent un plus grand nombre d'étiquettes cibles, et parfois se focaliser uniquement sur ces données cibles sera plus judicieux. En second lieu, la figure 6.7(b) reporte le comportement en fonction de κ (selon les meilleurs λ et β et avec $m^t = 10$). Comme attendu, un κ élevé, entre 0.9 et 0.99, est préféré : le gain est entre 0.1 et 0.5. Encore une fois, l'impact est plus significatif pour les tâches complexes. Pour les plus faciles, la table 6.3 montre que l'influence de κ sur la parcimonie est directe mais amoindrie lorsque B_R tend à être élevé. Enfin, comme espéré par le lemme 6.2, prendre en compte des étiquettes cibles implique des modèles moins parcimonieux.

Temps de calcul

Les temps de calculs moyens à paramètres fixés sont reportés sur la table 6.4. L'apprentissage d'un classifieur-SF correspond au temps de référence (le plus rapide avec SVM). On observe que les durées d'exécution de $\text{DASF}_{\text{it}_1}$ et celle de SF sont identiques. Le coût additionnel qu'implique les itérations de DASF reste raisonnale puisque pour au plus 10 itérations, il demande entre 4 et 8 fois plus de temps que $\text{DASF}_{\text{it}_1}$. De plus,



(a) *Les taux moyens pour chaque angle de rotation en fonction de la quantité d'étiquettes cibles considérées. Le résultat obtenu sans adaptation (i.e. sans étiquette source) est indiqué en gros pointillés noir.*

(b) *Les taux moyens en fonction de κ avec $m^t = 10$, $\lambda = 0.15$ et $\beta = 0.75$.*

FIGURE 6.7 – **Problème jouet.** *Les taux moyens de bonne classification obtenus en combinant des étiquettes sources et cibles avec SSDASF.*

Angle	20°	30°	40°	50°	60°	70°	80°	90°
$\kappa = 0$	24	22	16	18	11	8	17	8
$\kappa = 0.01$	18	19	11	17	15	7	15	6
$\kappa = 0.1$	22	18	13	10	11	9	4	11
$\kappa = 0.25$	21	19	14	17	10	17	6	11
$\kappa = 0.5$	19	18	19	11	12	8	13	10
$\kappa = 0.75$	19	17	16	20	9	14	8	14
$\kappa = 0.8$	**22**	20	10	16	12	14	9	16
$\kappa = 0.85$	24	18	24	18	7	14	13	11
$\kappa = 0.9$	24	21	18	23	14	12	9	10
$\kappa = 0.95$	21	**23**	**20**	**14**	**11**	12	16	14
$\kappa = 0.99$	7	7	7	7	12	**9**	**11**	**7**

TABLE 6.3 – **Problème jouet.** *Quantité moyenne de landmarks illustrant la parcimonie en fonction de κ avec $\lambda = 0.15$ et $\beta = 0.75$. En gras, les valeurs associées au modèle le plus performant.*

notre méthode DASF est trois fois plus rapide que DASVM, mais est cependant plus longue que TSVM (probablement dû à la recherche des couples $\mathcal{C}_{ST}$). Finalement, parmi les trois méthodes adaptatives — TSVM, DASVM et DASF — les plus faibles coût d'exécutions sont obtenus pour les problèmes de 40° à 70° de rotation. Nous n'avons pas reporté les coûts de SSDASF, car SSDASF et DASF ont un temps d'exécution du même ordre.

6.4.4 Classification d'images

Protocole

Dans cette section, nous expérimentons notre approche sur les corpora PascalVOC 2007 [Everingham *et al.*, 2007] et TrecVid 2007 [Smeaton *et al.*, 2009]. L'objectif est l'identification d'objets (concepts) visuels classiques dans des images. Le corpus TrecVid est constitué d'images extraites de vidéos et peut aussi être vu comme un corpus

Angle	20°	30°	40°	50°	60°*	70°*	80°*	90°*	Average
SVM	1	1	1	1	1	1	1	1	*1*
SF	1	1	1	1	1	1	1	1	*1*
TSVM	3	3	1	1	1	1	2	3	*1.8*
DASVM	29	14	13	8	8	19	26	28	*18.1*
DASF_{it_1}	1	1	1	1	1	1	1	1	*1*
DASF	8	7	6	6	4	6	7	6	*6.2*

Table 6.4 – **Problème jouet.** *Le temps de calcul moyen pour chaque méthode (SF est la mesure de base).*

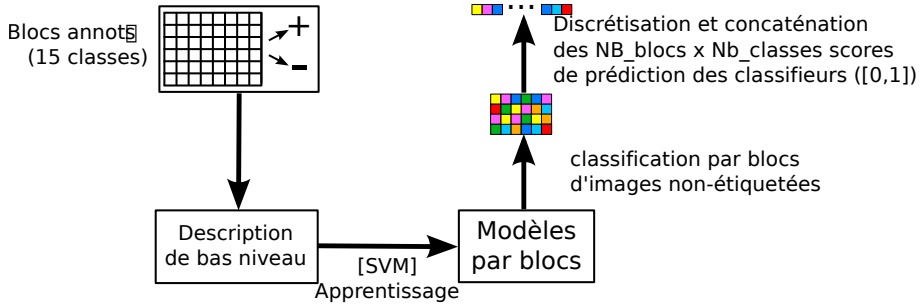


Figure 6.8 – **Classification d'images.** *Principe du descripteur visuel utilisé.*

d'image. Dans nos expériences, nous prenons, comme représentation des images un descripteur visuel défini par les scores de prédictions sur 15 concepts "intermédiaires" précédemment utilisés avec succès dans de précédentes évaluations TrecVid : (Animal, Building, Car, Cartoon, Explosion-Fire, Flag-US, Greenery, Maps, Road, Sea, Skin_face, Sky, Snow, Sports, Studio_setting). En utilisant le protocole de [Ayache *et al.*, 2007] illustré sur la figure 6.8, chacun d'entre eux est détecté par un classifieur-SVM à partir de moments de couleurs et d'histogrammes d'orientations sur 260 blocs de 32×32 pixels (la dimension est de 3900).

Adaptation lorsque le ratio d'étiquettes positives/négatives diffère entre les échantillons source et cible

Protocole Le corpus PascalVOC'07 est constitué d'un ensemble de 5 000 images d'apprentissage et de 5 000 de test. L'objectif est d'identifier 20 concepts. Les échantillons d'apprentissage et de test sont en fait relativement proches car ils sont issus du même domaine ($\hat{d}_{\mathcal{H}} \simeq 0.05$). L'adaptation de domaine n'est donc pas nécessaire. Cependant, nous voulons évaluer la capacité d'adaptation de notre algorithme lorsque le ratio d'images positives $(+/-)$ entre les échantillons d'apprentissage (source) et de test (cible) diffère, définissant ainsi une tâche d'adaptation plus difficile. Notre but n'est pas de proposer une solution à une telle situation (des méthodes spécifiques existent comme [Seah *et al.*, 2010]), mais plutôt d'évaluer si notre méthode peut améliorer

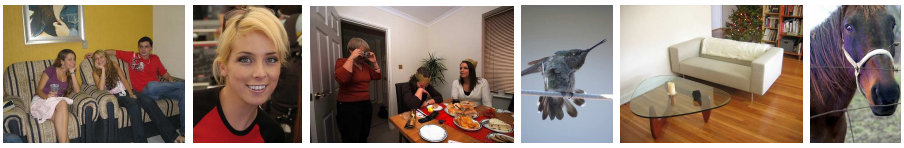


FIGURE 6.9 – **Classification d'images.** *Les 6 landmarks sélectionnés pour le concept* PERSON *(Pascal-VOC), les trois premières sont positives et les trois dernières négatives.*

les performances sur l'ensemble de test en évitant de faire du transfert négatif [11]. Puisque les deux domaines sont proches, nous observons uniquement les résultats de la méthode non supervisée DASF [12]. Dans ce jeu de données, le ratio $+/-$ est inférieur à 10%. Nous générons un échantillon source par concept, constitué de tous les exemples d'apprentissage positifs et d'exemples négatifs indépendemment tirés tel que le ratio $+/-$ soit $\frac{1}{3}/\frac{2}{3}$. L'échantillon cible est l'échantillon test originel. Pour les cinq algorithmes décrits précédemment, nous apprenons un classifieur binaire associé à chaque concept. Comme le ratio $+/-$ est faible sur l'échantillon de test, nous choisissons d'évaluer les performances au sens de la F-mesure définie par :

$$\frac{2 \times Precision \times Rappel}{Precision + Rappel}.$$

Résultats Les résultats sont reportés dans la table 6.5. Remarquons que TSVM et DASVM sont les moins performants, probablement à cause de l'absence supposée d'information sur le domaine cible et donc à l'impossibilité d'estimer le ratio $+/-$. Correctement paramétré, SVM obtient même de meilleures performances sur de nombreuses tâches (car les données d'apprentissage et de test sont issues du même domaine). DASF est, quant à lui, plus performant en moyenne et est le meilleur pour 18 concepts. Il améliore toujours le résultats d'un classifieur-SF et évite donc le transfert négatif. De plus, il construit toujours des modèles significativement plus parcimonieux. En illustration, la figure 6.9 est l'ensemble des *landmarks* sélectionnés pour le concept PERSON.

Temps de calcul

Le temps moyen d'exécution de chaque algorithme (à paramètres fixés) est reporté dans la table 6.6. Nous rappelons que la durée de base est celle de l'apprentissage d'un classifieur-SF. Pour ce corpus réel, DASVM est significativement plus coûteux. Contrairement au problème jouet, SVM est en moyenne plus long que SF et DASF est plus rapide que TSVM. En fait, les *landmarks* sont moins nombreux que les vecteurs de support. Encore une fois, une itération de notre algorithme s'éxécute dans le même temps que l'apprentissage d'un classifieur-SF. De plus, le coût additionel imposé par

11. Le transfert négatif signifie une perte de performance sur l'ensemble de test par rapport à un algorithme non adaptatif.

12. Dans le cas où les deux domaines sont identiques, considérer des étiquettes cibles revient à rajouter des étiquettes sources.

Conc.	bird	boat	bottle*	bus	car	cat	chair	cycle	cow	diningtable
SVM	0.18	0.29	0.01	0.16	0.28	**0.23**	0.24	0.10	0.15	0.15
SV	867	351	587	476	1096	882	1195	392	681	534
SF	0.18	0.27	0.11	0.12	0.34	0.20	0.21	0.10	0.11	0.10
Land.	237	203	233	212	185	**178**	241	139	239	**253**
TSVM	0.14	0.14	0.11	0.16	0.37	0.14	0.22	0.13	0.12	0.13
SV	814	704	718	445	631	779	864	390	888	515
DASVM	0.16	0.22	0.11	0.14	0.37	0.20	0.23	0.14	0.11	0.15
SV	922	223	295	421	866	1011	1418	706	335	536
DASF	**0.20**	**0.32**	**0.12**	**0.17**	**0.38**	**0.23**	**0.26**	**0.16**	**0.16**	**0.16**
Land.	**50**	**184**	**78**	**94**	**51**	378	**229**	**192**	**203**	372

Conc.	dog*	horse	monitor	motorbike	person*	plane	plant	sheep	sofa	train	Moy.
SVM	0.24	0.31	**0.16**	0.17	0.56	0.34	0.12	0.16	0.16	0.36	*0.22*
SV	436	761	698	670	951	428	428	261	631	510	*642*
SF	0.18	0.24	0.12	0.17	0.46	0.34	0.13	0.12	0.13	0.20	*0.19*
Land.	200	**247**	**203**	243	226	**178**	**236**	**128**	224	202	*210*
TSVM	0.22	0.17	0.12	0.12	0.44	0.18	0.10	0.12	0.15	0.19	*0.17*
SV	704	828	861	861	1111	585	406	474	866	652	*705*
DASVM	0.22	0.23	0.12	0.14	0.55	0.30	0.12	0.13	0.17	0.28	*0.20*
SV	**180**	802	668	841	303	356	1434	246	486	407	*622*
DASF	**0.25**	**0.32**	**0.16**	**0.18**	**0.58**	**0.35**	**0.15**	**0.20**	**0.18**	**0.42**	***0.25***
Land.	391	384	287	**239**	**6**	181	293	153	**167**	**75**	***200***

Table 6.5 – **Classification d'images.** *Les résultats (F-mesure) pour chaque concept (*Conc.*) sur l'échantillon cible/test de PascalVOC.* Moy. *correspond aux résultats en moyenne sur les* 20 *concepts.*

les itération de DASF reste raisonnable : il est en moyenne 5.4 fois plus lent que le temps d'une itération et est significativement plus rapide que DASVM et TSVM.

Adaptation de PascalVOC 2007 vers TrecVid 2007

Protocole Nous ne gardons plus que les six concepts communs à TrecVid'07 et PascalVOC'07. Pour chacun d'entre eux, nous considérons le même ensemble d'apprentissage issu de PascalVOC'07 et nous générons un échantillon cible/test constitué d'exemples issu de TrecVid avec le même ratio $+/-$. Dans cette situation, $\hat{d}_{\mathcal{H}}$ est de l'ordre de 1.4, les deux corpora sont très différents et la tâche d'adaptation de domaine est donc potentiellement complexe.

Résultats Les résultats évalués en terme de F-mesure sont reportés dans la table 6.7. DASF obtient en moyenne le meilleur résultat. Nous notons encore une fois que les modèles construits sont parcimonieux et performants. Enfin, pour ces tâches difficiles la similarité normalisée est toujours préférée (*). DASF est donc capable de s'améliorer à l'aide de similarités ni symétriques, ni semi-définies positives, telles que $K_{ST}(\cdot,\cdot)$ qui permet d'incorporer une information cible qui apparaît utile pour les tâches d'adaptation complexes.

Conc.	bird	boat	bottle*	bus	car	cat	chair	cycle	cow	diningtable
SVM	5.5	0.8	2.1	2	1.3	9	3.4	8.8	0.3	1.9
SF	1	1	1	1	1	1	1	1	1	1
TSVM	18.2	4.1	8	6	3.3	11.4	8.7	28.6	2.8	5.7
DASVM	4254	1440	3870	4860	1470	3428	2674	2828	900	1300
DASF_{it_1}	1	1	1	1	1	1	1	1	1	1
DASF	4.9	6	6	9	7.3	5	3.8	6.8	2.7	2.1

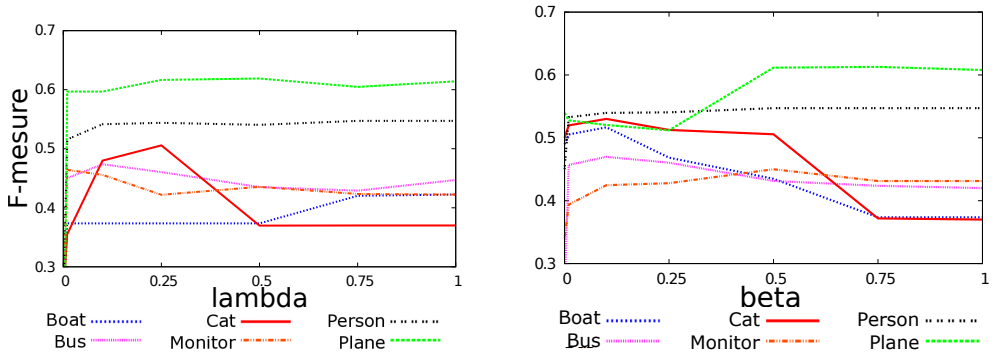
Conc.	dog*	horse	monitor	motorbike	person*	plane	plant	sheep	sofa	train	Moy.
SVM	1.4	3.7	4.2	5	2.1	2.9	0.5	1.4	2	1.4	*2.2*
SF	1	1	1	1	1	1	1	1	1	1	*1*
TSVM	1.9	7.4	12.2	5.9	4.5	10	0.7	4.4	15.9	10.1	*8.49*
DASVM	340	3553	4230	3060	675	1710	3900	1836	1912	1550	*2489*
DASF_{it_1}	1	1	1	1	1	1	1	1	1	1	*1*
DASF	6.1	5.8	8.8	5.5	5.6	5	3.7	3.6	5.7	5.5	*5.4*

Table 6.6 – **Classification d'images.** *Le temps moyen de calcul estimé comme un ratio du temps d'exécution de SF pour chaque méthode.*

Concept	boat*	bus*	car*	monitor*	person*	plane*	Moyenne
SVM	0.56	0.25	0.43	0.19	0.52	0.32	*0.38*
SV	351	476	1096	698	951	428	*667*
SF	0.49	0.46	0.50	0.34	0.45	0.54	*0.46*
Land.	214	224	**176**	246	226	178	*211*
TSVM	0.56	0.48	0.52	0.37	0.46	0.61	*0.50*
SV	498	535	631	741	1024	259	*615*
DASVM	0.52	0.46	**0.55**	0.30	0.54	0.52	*0.48*
SV	202	222	*627*	523	274	450	*383*
DASF	**0.57**	**0.49**	**0.55**	**0.42**	**0.57**	**0.66**	***0.54***
Land.	**120**	**130**	254	**151**	**19**	**7**	***113***

Table 6.7 – **Classification d'images.** *Les résultats (F-mesure) obtenus sur l'échantillon cible/test de TrecVid pour chaque concept.*

DASF : influence des hyperparamètres λ et β Dans ces expérimentations, nous détaillons sur la figure 6.10 l'impact de λ (selon le meilleur β) et de β (selon le meilleur λ). La figure 6.10(a) montre une influence relative de λ, excepté pour Car où la meilleure F-mesure est proche de 0.25 avec un gain d'au moins 0.15. Pour les autres concepts, le gain est inférieur à 0.1 et le λ le plus approprié, strictement positif, dépend de la tâche. À l'instar du problème jouet, la figure 6.10(b) indique clairement une influence de β majoritairement plus importante : pour Boat, Car et Plane le gain varie de 0.1 à 0.15 (pour Person, le gain est quasiment nul). Ainsi, en se focalisant plus finement sur β, la recherche des paramètres peut être allégée. La valeur de β, strictement positive, menant au meilleur classifieur appartient à $[0.01, 0.25]$, sauf pour Plane qui préfère une valeur supérieure à 0.5.



(a) *F-mesure en fonction de λ (β fixé).* **(b)** *F-mesure en fonction de β (λ fixé).*

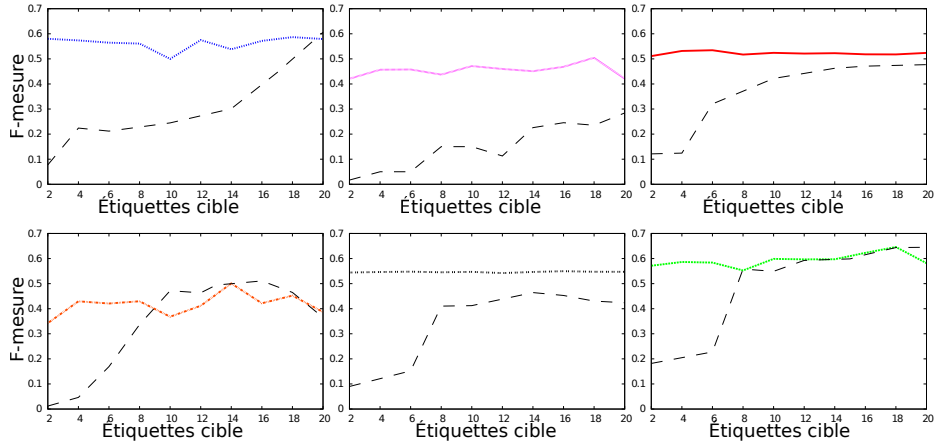
FIGURE 6.10 – **Classificatin d'images.** *Les résultats pour chaque concept en fonction de λ et β du problème* (6.2)*, avec DASF.*

CONCEPT	boat*	bus*	car*	monitor*	person*	plane*	MOYENNE
SVM	0.8	2	1.3	4.2	2.1	2.9	*2.2*
SF	1	1	1	1	1	1	*1*
TSVM	5	7	3.6	13.2	5.5	11	*7.5*
DASVM	3870	720	2370	2790	300	540	*1765*
DASF_{it_1}	1	1	1	1	1	1	*1*
DASF	9.5	10	8	23	3.4	10.5	*10.7*
SSDASF_{it_1}	4.4	2	0.4	1	0.6	1.5	*1.6*
SSDASF	29.8	14.4	6.6	29	3.4	13	*16*

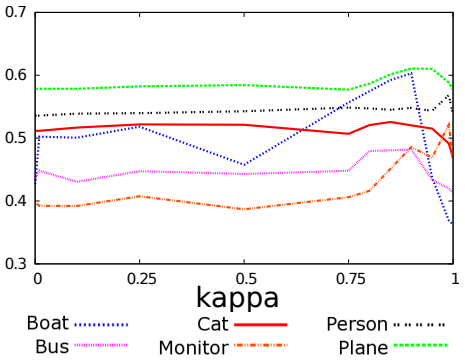
TABLE 6.8 – **Classification d'images.** *Le temps d'exécution moyen de chaque méthode mesuré comme un ratio du temps d'exécution de SF.*

SSDASF : Influence de données étiquetées cibles Tout d'abord, sur la figure 6.11(a) nous détaillons les résultats moyens obtenus pour chaque concept. Ils améliorent ceux sans étiquette cible (avec $\kappa = 0$). De plus, lorsque le nombre d'étiquettes cibles est strictement inférieur à 8, les modèles inférés sont toujours plus performants que le classifieur-SF appris uniquement à partir des étiquettes cibles. Plus de 20 exemples sont parfois nécessaires pour atteindre les performances de SSDASF, montrant l'utilité des étiquettes cibles. Ensuite, nous avons testé différentes valeurs de κ (selon les meilleurs λ et β et avec $m^t = 10$) et reporté les résultats sur la figure 6.11(b). Les valeurs de κ les plus pertinentes se situent entre 0.9 et 0.99. Pour ce jeu de données plus complexe à traiter, les données cibles apportent donc une information importante. Finalement, le descripteur utilisé ne semble pas être assez expressif, expliquant la difficulté à obtenir de meilleurs résultats. En effet, en traitement d'images ou en multimédia, à l'image de la section 4.3 du chapitre 4, les données sont souvent représentées en fonction de plusieurs modalités pour permettre une plus grande expressivité, une perspective serait alors d'adapter SSDASF à la multi-modalité.

Temps de calcul La table 6.8 correspond aux temps de calcul moyens (à paramètres fixés) définis comme un ratio du temps d'exécution de l'apprentissage d'un classifieur-



(a) *F-mesure en fonction de la quantité de la quantité d'étiquettes cibles considérées. Le résultat obtenu sans adaptation (c'est-à-dire sans étiquette source) est indiqué en gros pointillés noirs.*



(b) *F-mesure en fonction de* κ *avec* $m^t = 10$, $\lambda = 0$ *et* β *fixés*

FIGURE 6.11 – **Classification d'images.** *F-mesure obtenue pour chaque concept en combinant des étiquettes sources et cibles dans le problème* (6.7)*, avec SSDASF.*

SF. L'unique différence avec la précédente tâche (voir la section 6.4.4) est que TSVM est plus rapide que DASF, probablement dû à la difficulté de la tâche. En fait, dans la section 6.4.4 la divergence entre les distributions marginales était très faible, alors qu'ici la divergence est élevée. La différence élevée entre les deux domaine implique une recherche des couples plus complexe : lorsque deux points sont loin, la minimisation de la fonction objectif du problème (6.5) nécessite plus de temps.

Enfin, SSDASF (et sa première itération) s'exécute 1.5 plus lentement que la version non supervisée DASF. En fait, la définition du problème (6.7) se fait en ajoutant des contraintes à (6.2), impliquant une exécution plus longue. Que ce soit pour DASF ou pour SSDASF, le coût supplémentaire dû aux itérations est relativement raisonnable avec un facteur de 10 en moyenne. L'approche itérative reste donc compétitive en terme de temps de calcul.

6.5 Synthèse

Dans ce chapitre nous avons proposé un nouvel algorithme d'adaptation de domaine utilisant les outils offerts par la théorie de [Balcan *et al.*, 2008a, Balcan *et al.*, 2008b]. Cette théorie permet d'apprendre un classifieur linéaire dans un espace de projection explicite défini à partir d'une fonction de similarité (ϵ,γ,τ)-bonne, potentiellement non symétrique et non semi-définie positive. Avec une perspective d'adaptation de domaine, notre méthode régularise cet apprentissage de sorte que l'espace de projection induit permette un rapprochement des deux domaines tout en gardant de bonnes performances sur le domaine source. Cette minimisation jointe est un clair avantage à notre méthode. D'un point de vue pratique, le processus itératif permet d'alléger la recherche de l'espace de projection en pondérant les similarités, mais, couplé au réglage des hyperparamètres. Il rend la procédure assez lourde à mettre en œuvre. De plus, le contexte des fonctions de similarité (ϵ,γ,τ)-bonnes implique une pré-sélection de *landmarks* assez représentatif des données à traiter et reste une problématique ouverte. Nous pourrions exploiter les récents travaux de [Gong *et al.*, 2013] pour proposer une sélection judicieuse de ces points dans notre contexte. D'un point de vue théorique, nous avons analysé la capacité en généralisation de l'algorithme mettant en avant l'importance de la régularisation en démontrant sa robustesse. Une telle analyse reste cependant assez complexe et nous aimerions étudier plus finement notre approche en s'inpirant, par exemple, du λ-shift proposé après le développement de notre approche (présenté en section 2.1.4 du chapitre 2). En outre, nous avons étendu notre méthode à l'utilisation de quelques étiquettes cibles. Les approches ont démontré empiriquement de bonnes capacités d'adaptation sur des tâches variées. Cependant, cet algorithme révèle d'autres inconvénients. En effet, notre méthode se base sur la théorie classique de l'adaptation de domaine qui peut être vue comme une analyse en pire cas puisque la divergence $\frac{1}{2}d_{\mathcal{H}}(D_S, D_T)$ entre les domaines correspond au désaccord maximal entre deux classifieurs. De plus, le processus de minimisation de la divergence que nous avons suivi se base sur une heuristique puisque minimiser $\frac{1}{2}d_{\mathcal{H}}(S_u, T_u)$ et l'erreur source coinjointement est complexe (voir le chapitre 2). Il n'y a donc aucune justification théorique de sa minimisation conjointe avec le risque source. Enfin, cet algorithme est spécifique à la combinaison de fonctions de similarité (ϵ,γ,τ)-bonnes. Pour contrer ces inconvénients nous apportons dans le chapitre suivant une nouvelle vision de l'adaptation de domaine en faisant appel à la théorie PAC-Bayésienne permettant de s'attaquer aux défauts liés à (SS)DASF. Cette vision originale implique une analyse en moyenne plus fine, mais aussi une jutification théorique à la minimisation directe de la divergence et de l'erreur source.

7 Analyse PAC-Bayésienne de l'adaptation de domaine

Comme souligné dans la synthèse du chapitre précédent, l'analyse classique du problème de l'adaptation de domaine souffre d'un inconvénient majeur : la minimisation de la $\mathcal{H}$-divergence, qui est une mesure dans le pire cas, rend difficile une optimisation conjointe de la divergence et de l'erreur sur le domaine source. Pour contrer ce problème, nous étudions l'adaptation de domaine avec un point de vue PAC-Bayésien. D'une part, nous savons que la théorie PAC-Bayésienne permet de proposer des analyses en moyenne du processus d'apprentissage et induit, en ce sens, des bornes en généralisation plus précises. D'autre part, nous rappelons qu'elle se focalise sur l'apprentissage de votes de majorité qui est au cœur de ce manuscrit. Nous proposons donc, dans ce chapitre, la première analyse PAC-Bayésienne du problème de l'adaptation de domaine. Notre contribution repose sur trois points principaux. Premièrement, nous définissons en section 7.1.2 une divergence entre les domaines appropriée à l'analyse PAC-Bayésienne, bien plus simple à estimer et à minimiser, et dont le processus de minimisation empirique est consistant. À partir de

cette mesure, nous démontrons en section 7.1.5 une borne d'adaptation de domaine que nous analysons selon les trois types de bornes PAC-Bayésiennes classiques pour le classifieur de Gibbs associé au vote de majorité ρ-pondéré. Nous en dérivons, dans la section 7.2, un premier algorithme PAC-Bayésien pour l'adaptation de domaine en spécialisant ce résultat aux classifieurs linéaires. Cet algorithme montre, en section 7.3, des comportements empiriques très prometteurs sur un jeu de données synthétique et une tâche d'analyse d'avis.

Ce chapitre a été publié dans les conférences ICML 2013 [Germain *et al.*, 2013a] et CAp 2013 [Germain *et al.*, 2013b]. Il a de plus donné lieu à une communication scientifique non publiée au *workshop* de NIPS 2012 *Multi-Trade-offs in Machine Learning* [1].

7.1 Bornes d'adaptation de domaine pour le classifieur de Gibbs

7.1.1 Notations

Dans ce chapitre, nous nous plaçons uniquement dans le cadre de l'adaptation de domaine non supervisée (sans étiquette cible) d'un domaine source P_S vers un domaine cible P_T. Nous ne considérons que des problèmes de classification binaire où $X \in \mathbb{R}^d$ est l'espace de description et $Y = \{-1; +1\}$ est l'ensemble des étiquettes possibles. Nous gardons les notations du chapitre précédent, à l'exception près que $\mathcal{H}$ est un ensemble de classifieurs de X vers Y. Notons que les résultats théoriques présentés dans ce chapitre sont valides pour des votants à valeurs réelles et des fonctions de pertes vérifiant l'inégalité triangulaire.

Nous reprenons l'approche PAC-Bayésienne afin d'étudier les capacités d'adaptation de votes de majorité ρ-pondéré $B_\rho(\cdot)$ sur l'ensemble des votants issus de $\mathcal{H}$:

$$\forall \mathbf{x} \in X,\ B_\rho(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right].$$

Nous rappelons qu'étant donnée une distribution prior π sur $\mathcal{H}$, le but de l'apprenant PAC-Bayésien est de trouver une distribution posterior ρ sur $\mathcal{H}$ amenant à de bonnes garanties en généralisation pour $B_\rho(\cdot)$. Dans le chapitre 3, nous avons vu que ces garanties se concentrent sur l'erreur du classifieur stochastique de Gibbs $G_\rho(\cdot)$ associé à ρ, qui étiquette un exemple $\mathbf{x}$ en tirant aléatoirement un votant dans $\mathcal{H}$ selon ρ puis en retournant $h(\mathbf{x})$ (ou $\operatorname{sign}[h(\mathbf{x})]$ si les votants sont à valeurs réelles). Étant donné un domaine P sur $X \times Y$, l'erreur du classifieur de Gibbs est alors vue comme l'espérance des erreurs des votants de $\mathcal{H}$ selon ρ :

$$\mathbf{R}_P(G_\rho) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_P(h).$$

Afin d'analyser le problème de l'adaptation de domaine sous l'angle de la théorie PAC-Bayésienne, nous devons donc étudier le comportement du moyennage des erreurs

1. https://sites.google.com/site/multitradeoffs2012/

selon ρ. En outre, la dérivation d'une borne en adaptation de domaine non supervisée requiert l'utilisation cruciale d'une mesure de divergence entre les distributions marginales selon X associées aux domaines (voir section 2.2.1, chapitre 2). Ainsi, nous définissons dans la section suivante une divergence entre distributions facilement estimable à partir des échantillons non étiquetés S_u et T_u et s'exprimant comme une espérance sur $\mathcal{H}$ selon la distribution ρ.

7.1.2 Le ρ-désaccord : une divergence appropriée à l'analyse PAC-Bayésienne

Notre divergence se définit comme une pseudo-métrique [2] mesurant l'écart entre les désaccords mesurés sur les domaines. Elle quantifie la différence structurelle entre les marginales D_S et D_T en terme de posterior ρ sur l'ensemble $\mathcal{H}$. Puisqu'en analyse PAC-Bayésienne, le but est d'apprendre un vote de majorité ρ-pondéré dont l'erreur $\mathbf{R}_P(B_\rho)$ est la plus faible possible, nous proposons de suivre l'idée portée par la C-borne du théorème 3.1 dont nous rappelons la forme :

$$\mathbf{R}_P(B_\rho) \leq 1 - \frac{\big(1 - 2\mathbf{R}_P(G_\rho)\big)^2}{1 - 2\mathbf{R}_D(G_\rho, G_\rho)} \,,$$

où $\mathbf{R}_D(G_\rho, G_\rho) = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{I}\left(h(\mathbf{x}) \neq h'(\mathbf{x})\right) = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_D(h, h')$ est l'espérance selon ρ des désaccords entre les classifieurs de $\mathcal{H}$.
Concrètement, étant donnés les domaines P_S et P_T sur $X \times Y$ et une distribution ρ sur $\mathcal{H}$, si $\mathbf{R}_{P_S}(G_\rho)$ et $\mathbf{R}_{P_T}(G_\rho)$ sont similaires, alors $\mathbf{R}_{P_S}(B_\rho)$ et $\mathbf{R}_{P_T}(B_\rho)$ sont similaires lorsque $\mathbf{E}_{(h,h')\sim\rho^2} \mathbf{R}_{D_S}(h, h')$ et $\mathbf{E}_{(h,h')\sim\rho^2} \mathbf{R}_{D_T}(h, h')$ sont proches. Ainsi, les deux domaines P_S et P_T sont proches selon ρ si l'écart entre $\mathbf{E}_{(h,h')\sim\rho^2} \mathbf{R}_{D_S}(h, h')$ et $\mathbf{E}_{(h,h')\sim\rho^2} \mathbf{R}_{D_T}(h, h')$ tend à être faible. Nous nommons notre divergence le ρ-désaccord et nous la définissons en fonction de ces deux désaccords comme suit :

Définition 7.1 *Soit $\mathcal{H}$ un ensemble de classifieurs de X vers Y. Pour toutes distributions marginales D_S et D_T sur X, et pour toute distribution ρ sur $\mathcal{H}$, le ρ-désaccord entre les domaines* $\mathrm{dis}_\rho(D_S, D_T)$ *entre D_S et D_T est défini par :*

$$\mathrm{dis}_\rho(D_S, D_T) = \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{D_T}(h, h') - \mathbf{R}_{D_S}(h, h')\right] \right| .$$

Étant donnés deux échantillons non étiquetés S_u et T_u dont les éléments sont respectivement tirés selon D_S et D_T, le ρ-désaccord empirique $\mathrm{dis}_\rho(S_u, T_u)$ *calculé sur S_u et T_u est défini par :*

$$\mathrm{dis}_\rho(S_u, T_u) = \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{T_u}(h, h') - \mathbf{R}_{S_u}(h, h')\right] \right| .$$

Notons qu'il est trivial de démontrer que le ρ-désaccord $\mathrm{dis}_\rho(\cdot, \cdot)$ est symétrique et vérifie l'inégalité triangulaire. Les résultats de la section suivante justifient de la consistance du processus de minimisation empirique du ρ-désaccord. En effet, la valeur

2. Contrairement à une métrique, deux points ne sont pas nécessairement discernables par la pseudo-métrique : il est possible d'avoir $d(x, x') = 0$ pour des valeurs distinctes $x \neq x'$.

réelle $\text{dis}_\rho(D_S, D_T)$ peut être bornée par les quantités classiques de la théorie PAC-Bayésienne : le ρ-désaccord empirique $\text{dis}_\rho(S_u, T_u)$ estimé à partir des échantillons source S_u et cible T_u (non étiquetés) et la KL-divergence entre les distributions prior et posterior sur $\mathcal{H}$.

7.1.3 Consistance de la minimisation empirique du ρ-désaccord

Nous présentons, ci-dessous, les bornes en généralisation PAC-Bayésiennes pour notre ρ-désaccord entre distributions sous les trois formes "classiques" présentées en section 3.2.3 du chapitre 3. Pour rappel, ces trois formes ont leur propre avantage. La version de McAllester du corollaire 3.2 est la plus simple a interpréter, celle de Langford-Seeger du corollaire 3.1 plus précise que celle de McAllester et celle de Catoni du corollaire 3.3 qui est la plus intéressante d'un point de vue algorithmique car elle justifie du contrôle du compris entre la complexité et l'erreur empirique. Le lecteur peut prendre à sa guise la borne avec laquelle il est le plus à l'aise.

Tout d'abord, le théorème suivant énonce la version de la borne "à la McAllester" plus simple à interpréter.

Théorème 7.1 *Pour toutes distributions marginales D_S et D_T sur X, pour tout ensemble de classifieurs $\mathcal{H}$, pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0, 1]$, avec une probabilité d'au moins $1 - \delta$ sur le choix aléatoire des échantillons $S_u \sim (D_S)^{m_u}$ et $T_u \sim (D_T)^{m_u}$, pour toute distribution ρ sur $\mathcal{H}$, on a :*

$$\left|\text{dis}_\rho(D_S, D_T) - \text{dis}_\rho(S_u, T_u)\right| \leq 2\sqrt{\frac{2}{m_u}\left[2\,\text{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m_u}}{\delta}\right]}. \tag{7.1}$$

Démonstration. En annexe F.1. □

Ce résultat confirme de la consistance de la minimisation empirique du ρ-désaccord. Notons que ce résultat reste valide lorsque la taille des échantillons S_u et T_u diffèrent. Si m_u^s est la taille de S_u et m_u^t celle de T_u, la borne devient :

$$\begin{aligned}\left|\text{dis}_\rho(D_S, D_T) - \text{dis}_\rho(S_u, T_u)\right| \leq &\sqrt{\frac{2}{m_u^s}\left[2\,\text{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m_u^s}}{\delta}\right]} \\ &+ \sqrt{\frac{2}{m_u^t}\left[2\,\text{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m_u^t}}{\delta}\right]}.\end{aligned}$$

Il est plus simple de donner du sens à ce type de borne, puisqu'elle majore directement l'écart entre la valeur réelle du ρ-désaccord et son estimation empirique. Cependant, nous rappelons que cette approche n'offre pas les meilleures garanties. L'approche "à la Langford-Seeger", comparant les valeurs réelle et empirique à l'aide de la fonction $\text{kl}(\cdot, \cdot)$, permet d'obtenir un résultat plus précis. Par souci de complétude, nous énonçons ce résultat.

Théorème 7.2 *Pour toutes distributions marginales D_S et D_T sur X, pour tout ensemble de classifieurs $\mathcal{H}$ de X vers Y, pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0, 1]$, avec une*

probabilité d'au moins $1-\delta$ *sur le choix de* $S_u \times T_u \sim (D_S \times D_T)^{m_u}$, *pour toute distribution* ρ *sur* $\mathcal{H}$, *on a :*

$$\mathrm{kl}\left(\frac{\mathrm{dis}_\rho(S_u,T_u)+1}{2}\middle\|\frac{\mathrm{dis}_\rho(D_S,D_T)+1}{2}\right) \leq \frac{1}{m_u}\left[2\,\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m_u}}{\delta}\right].$$

Démonstration. En annexe F.2. □

On voit clairement que cette borne, dont le taux de convergence est de $O\left(\frac{1}{m_u}\right)$, est plus précise que celle du théorème 7.1 qui varie en $O\left(\sqrt{\frac{2}{m_u}}\right)$, mais est plus difficilement interprétable à cause du terme lié à la fonction $\mathrm{kl}(\cdot,\cdot)$.

Finalement, puisque nous avons l'objectif de proposer un premier algorithme PAC-Bayésien pour l'adaptation de domaine, nous montrons la consistance de la minimisation empirique du ρ-désaccord à l'aide de l'approche "à la Catoni". Nous rappelons, en effet, que son approche permet de contrôler, via un paramètre, le compromis estimateur et KL-divergence.

Théorème 7.3 *Pour toutes distributions marginales* D_S *et* D_T *sur* X, *pour tout ensemble de classifieur* $\mathcal{H}$ *de* X *vers* Y, *pour toute distribution prior* π *sur* $\mathcal{H}$, *pour tout réel* $A > 0$ *et pour tout* $\delta \in (0,1]$, *avec une probabilité d'au moins* $1-\delta$ *sur le choix de* $S_u \times T_u \sim (D_S \times D_T)^{m_u}$, *pour toute distribution* ρ *sur* $\mathcal{H}$, *on a :*

$$\mathrm{dis}_\rho(D_S,D_T) \leq \frac{2A}{1-e^{-2A}}\left[\mathrm{dis}_\rho(S_u,T_u)+\frac{2\,\mathrm{KL}(\rho\|\pi)+\ln\frac{2}{\delta}}{Am_u}+1\right]-1.$$

Démonstration. En annexe F.3. □

De manière similaire au corollaire 3.3, la borne est consistante lorsque $A=\frac{1}{2\sqrt{m_u}}$. Elle tend vers $1\times\left[\mathrm{dis}_\rho(S_u,T_u)+0+1\right]-1$ lorsque m_u tend vers $+\infty$.

Avant de dériver, en section 7.1.5, une nouvelle borne d'adaptation de domaine spécifique au classifieur de Gibbs, et donc à la théorie PAC-Bayésienne, nous comparons notre mesure de ρ-désaccord entre distributions avec la $\mathcal{H}$-divergence [Ben-David *et al.*, 2010, Ben-David *et al.*, 2007] définie dans la définition 2.4 de la section 2.2.2.

7.1.4 Comparaison de la $\mathcal{H}$-divergence et du ρ-désaccord

Rappelons avant tout, la formulation de la $\mathcal{H}$-divergence $\frac{1}{2}d_{\mathcal{H}}(\cdot,\cdot)$ équivalente à la *discrepancy* $\mathrm{disc}_{\ell_{0\text{-}1}}(\cdot,\cdot)$ de [Mansour *et al.*, 2009a] associée à la fonction de perte $0-1$:

$$\begin{aligned}\tfrac{1}{2}d_{\mathcal{H}}(D_S,D_T) &= \mathrm{disc}_{\ell_{0\text{-}1}}(D_S,D_T)\\ &= \sup_{(h,h')\in\mathcal{H}^2}\left|\mathbf{R}_{D_T}(h,h')-\mathbf{R}_{D_S}(h,h')\right|.\end{aligned}$$

D'une part, la minimisation de la $\mathcal{H}$-divergence entre deux distributions n'est pas une tâche facile lorsque l'on veut simultanément minimiser l'erreur source. D'autre part,

elle revient à trouver la paire de classifieurs dans $\mathcal{H}$ maximisant le désaccord. Elle correspond donc à une divergence dans le pire cas qui admet la même valeur quel que soit le classifieur de $\mathcal{H}$ considéré : cette divergence ne permet pas d'étudier le classifieur sur lequel on veut apporter des garanties théoriques.
En s'exprimant comme une espérance des désaccords, notre mesure de ρ-désaccord offre deux avantages. Le premier est qu'elle s'avère beaucoup plus simple à estimer et à minimiser puisqu'il suffit de calculer une moyenne en fonction de ρ : son optimisation peut directement se réaliser en minimisant son estimation empirique $\mathrm{dis}_\rho(S_u, T_u)$ et la KL-divergence. Le second est que sa définition dépend de la distribution ρ considérée : elle se montre adaptée au classifieur que l'on apprend. En outre, il est facile de montrer que pour toutes distributions D_S et D_T, $\mathrm{dis}_\rho(D_S, D_T)$, qui calcule une espérance, est plus faible que $\frac{1}{2}d_{\mathcal{H}}(D_S, D_T)$, qui renvoie la valeur maximale. Ceci rend nos résultats plus précis. En fait, pour tout ensemble de classifieurs $\mathcal{H}$ de X vers Y, pour toutes distributions D_S et D_T sur X et ρ sur $\mathcal{H}$, on a :

$$\begin{aligned}
\tfrac{1}{2}d_{\mathcal{H}}(D_S, D_T) &= \sup_{(h,h')\in\mathcal{H}^2} \left|\mathbf{R}_{D_T}(h,h') - \mathbf{R}_{D_S}(h,h')\right| \\
&\geq \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left|\mathbf{R}_{D_T}(h,h') - \mathbf{R}_{D_S}(h,h')\right| \\
&\geq \left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{D_T}(h,h') - \mathbf{R}_{D_S}(h,h')\right]\right| \\
&= \mathrm{dis}_\rho(D_S, D_T).
\end{aligned}$$

Nous voyons maintenant comment cette nouvelle divergence nous permet de dériver des garanties PAC-Bayésienne pour l'adaptation de domaine.

7.1.5 L'analyse PAC-Bayésienne de l'adaptation de domaine

Le théorème suivant énonce la borne d'adaptation de domaine appropriée à l'analyse PAC-Bayésienne et qui est le résultat principal de ce chapitre.

Théorème 7.4 *Soit P_S et P_T deux domaines sur $X \times Y$ dont D_S et D_T sont les marginales respectives sur X. Soit $\mathcal{H}$ un ensemble de classifieurs de X vers Y, alors pour toute distribution ρ sur $\mathcal{H}$, on a :*

$$\mathbf{R}_{P_T}(G_\rho) \leq \mathbf{R}_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S, D_T) + v_\rho, \tag{7.2}$$

où $v_\rho = \mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*})$,
avec $\rho_T^* = \mathrm{argmin}_\rho\, \mathbf{R}_{P_T}(G_\rho)$ la distribution posterior optimale sur le domaine cible P_T.

Démonstration. Soit $\mathcal{H}$ un ensemble de classifieurs de X vers Y. Soit ρ une distribution sur $\mathcal{H}$. Soit $\rho_T^* = \mathrm{argmin}_\rho\, \mathbf{R}_{P_T}(G_\rho)$ la distribution posterior optimale sur P_T. L'inégalité triangulaire et le fait que pour tout h issu de $\mathcal{H}$ et pour toute marginale D on ait :

$$\mathbf{R}_D(G_\rho, h) = \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{h'\sim\rho} \mathbf{I}[h'(\mathbf{x}) \neq h(\mathbf{x})],$$

nous permettent d'écrire :

$$\begin{aligned}
\mathbf{R}_{P_T}(G_\rho) &\leq \mathop{\mathbf{E}}_{h\sim\rho}\Big(\mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_{\rho_T^*}, G_\rho) + \mathbf{R}_{D_T}(G_\rho, h)\Big)\\
&\leq \mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_{\rho_T^*}, G_\rho) + \mathop{\mathbf{E}}_{h\sim\rho}\Big(\mathbf{R}_{D_T}(G_\rho, h) - \mathbf{R}_{D_S}(G_\rho, h) + \mathbf{R}_{D_S}(G_\rho, h)\Big)\\
&\leq \mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{D_S}(G_\rho, h)\\
&\quad + \left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left[\mathbf{R}_{D_T}(h,h') - \mathbf{R}_{D_S}(h,h')\right]\right|\\
&\leq \mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*}) + \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{P_S}(h)\\
&\quad + \left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left[\mathbf{R}_{D_T}(h,h') - \mathbf{R}_{D_S}(h,h')\right]\right| \qquad (7.3)\\
&= \mathbf{R}_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S, D_T) + \mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*}).
\end{aligned}$$

L'inégalité (7.3) provient des inégalités :

$$\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{D_S}(G_\rho, h) \leq \mathbf{R}_{P_S}(G_\rho) + \mathbf{R}_{P_S}(G_\rho)$$
$$\text{et : } \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{D_S}(G_\rho, h) \leq \mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_{\rho_T^*}, G_\rho).$$

Ainsi, on borne :

$$\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{D_S}(G_\rho, h) = \frac{1}{2}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{D_S}(G_\rho, h) + \frac{1}{2}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{D_S}(G_\rho, h)$$
$$\mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*}) + \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_{P_S}(h).$$

Finalement, on obtient la borne (7.2) en posant :

$$v_\rho = \mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*}).$$

□

En général, cette borne ne peut-être comparée aux analyses classiques de [Ben-David *et al.*, 2010] ou de [Mansour *et al.*, 2009a] présentées en section 2.2.3 du chapitre 2. Nous pouvons néanmoins souligner les deux points suivants, liés au fait que notre borne s'exprime comme un compromis entre différentes quantités.

- $\mathbf{R}_{P_S}(G_\rho)$ et $\mathrm{dis}_\rho(D_S, D_T)$ sont semblables aux deux premiers termes de la borne d'adaptation de domaine[3] du théorème 2.3 [Ben-David *et al.*, 2010] : $\mathbf{R}_{P_S}(G_\rho)$ est l'erreur source du classifieur de Gibbs associé à la distribution ρ étudiée et $\mathrm{dis}_\rho(D_T, D_S)$ mesure la divergence entre D_S et D_T mais dépend du classifieur final considéré.

- Similairement au dernier terme[4] de la borne d'adaptation de domaine du théorème 2.5 [Mansour *et al.*, 2009a], le terme $v_\rho = \mathbf{R}_{P_T}(G_{\rho_T^*}) + \mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) +$

3. Nous rappelons la forme de la borne d'adaptation de [Ben-David *et al.*, 2010], pour tout $h \in \mathcal{H}$, on a : $\mathbf{R}_{P_T}(h) \leq \mathbf{R}_{P_S}(h) + \frac{1}{2}d_{\mathcal{H}}(D_S, D_T) + \nu$.

4. Nous rappelons la forme du dernier terme de la borne d'adaptation de [Mansour *et al.*, 2009a] : $v = \mathbf{R}_{P_T}(h_T^*) + \mathbf{R}_{D_S}(h_S^*, h_T^*)$, où h_T^*, respectivement h_S^*, est le meilleur classifieur cible, respectivement source, dans $\mathcal{H}$.

$\mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*})$ dépend directement du classifieur optimal, c'est-à-dire de la distribution posterior ρ_T^*, que l'on peut obtenir sur le domaine cible. Notons, cependant, qu'il ne dépend pas de la solution optimale source, rendant le résultat plus pertinent, mais, d'un point de vue numérique, implique un terme supplémentaire. Plus précisément, le terme $\mathbf{R}_{P_T}(G_{\rho_T^*})$ correspond à la plus faible erreur que l'on peut espérer obtenir sur le domaine cible. Les deux autres termes $\mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*})$ et $\mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*})$ dépendent de ρ et quantifient à quel point la distribution ρ est proche (en terme de désaccords) du classifieur de Gibbs optimal, à la fois sur le domaine cible et le domaine source.

D'après le résultat du théorème 7.4, l'adaptation est donc possible si la distribution optimale ρ_T^* admet un erreur faible sur le domaine cible (c'est une hypothèse classique). De plus, la quantité $\mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*})$, qui peut être vue comme une mesure de la capacité d'adaptation en terme d'étiquetage, doit être faible : G_ρ doit être en accord — sur les deux domaines — avec la solution optimale $G_{\rho_T^*}$, autrement dit les deux domaines doivent être reliés [5].

Finalement, le théorème 7.4 conduit aux trois bornes PAC-Bayésiennes suivantes portant à la fois sur l'erreur source empirique du classifieur de Gibbs et la pseudo-métrique de ρ-désaccord estimée sur des échantillons source et cible. Tout d'abord, nous énonçons la version "à la McAllester".

Théorème 7.5 *Pour tous domaines P_S et P_T (de marginales respectives D_S et D_T) sur $X \times Y$, pour tout espace d'hypothèses $\mathcal{H}$, pour toute distribution prior π sur $\mathcal{H}$, pour tout $\delta \in (0, 1]$, avec une probabilité de $1 - \delta$ sur le choix des échantillons aléatoires $S \sim (P_S)^{m^s}$, $S_u \sim (D_S)^{m_u}$, et $T_u \sim (D_T)^{m_u}$, pour toute distribution ρ sur $\mathcal{H}$, on a :*

$$\mathbf{R}_{P_T}(G_\rho) \leq \mathbf{R}_S(G_\rho) + \mathrm{dis}_\rho(S_u, T_u) + \sqrt{\frac{2}{m^s}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m^s}}{\delta}\right]} + 2\sqrt{\frac{2}{m_u}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{8\sqrt{m_u}}{\delta}\right]} + v_\rho. \tag{7.4}$$

Démonstration. Le résultat est directement obtenu en combinant le corollaire 3.2 avec les théorèmes 7.1 et 7.5. □

Notons que si tous les échantillons sont de même taille, c'est-à-dire si $m_u = m^s = m$, la borne (7.4) devient :

$$\mathbf{R}_{P_T}(G_\rho) \leq \mathbf{R}_S(G_\rho) + \mathrm{dis}_\rho(S_u, T_u) + \sqrt{\frac{18}{m}\left[\mathrm{KL}(\rho\|\pi) + 4\ln\frac{11\sqrt{m}}{\delta}\right]} + v_\rho.$$

Supposons les hypothèses suivantes :

5. Pour faciliter l'interprétation, en posant ρ_S^* comme la distribution optimale source, on peut borner :

$$\mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_\rho, G_{\rho_T^*}) \leq \mathbf{R}_{D_T}(G_\rho, G_{\rho_T^*}) + \mathbf{R}_{D_S}(G_\rho, G_{\rho_S^*}) + \mathbf{R}_{D_S}(G_{\rho_S^*}, G_{\rho_T^*}).$$

(i) il existe une distribution posterior amenant à une erreur cible faible (l'hypothèse classique) ;

(ii) il existe un lien entre les domaines en termes d'accord d'étiquetage pour toute distribution ρ, autrement dit qu'un ρ-désaccord $\mathrm{dis}_\rho(D_S, D_T)$ faible implique un v_ρ négligeable.

Alors, puisque notre mesure de ρ-désaccord est facilement minimisable et dépend directement de la distribution ρ, une solution naturelle pour définir un algorithme PAC-Bayésien d'adaptation de domaine non supervisée est l'optimisation de la borne du théorème 7.5 en négligeant [6] le terme v_ρ. Puisque le ρ-désaccord est simple à minimiser, un avantage majeur de ce résultat est qu'il justifie théoriquement de l'optimisation simultanée de l'erreur source et de la divergence entre les distributions marginales. Cette propriété n'était pas aussi directe dans le cas des cadres classiques présentés dans le chapitre 2 et utilisés dans le chapitre précédent.

Avant de présenter le théorème dans sa version "à la Catoni" qui va nous permettre de dériver facilement un algorithme, nous énonçons la version "à la Langford-Seeger".

Théorème 7.6 *Pour tous domaines P_S et P_T sur $X \times Y$ (de marginales respectives D_S et D_T), pour tout ensemble de classifieurs $\mathcal{H}$ de X vers Y, pour toute distribution prior π sur $\mathcal{H}$ et pour tout $\delta \in (0, 1]$, avec une probabilité de $1 - \delta$ sur le choix aléatoire d'échantillons de même tailles $S \sim (P_S)^m$, $S_u \sim (D_S)^m$ et $T_u \sim (D_T)^m$, pour toute distribution ρ sur $\mathcal{H}$, on a :*

$$\mathbf{R}_{P_T}(G_\rho) \leq \sup R_\rho + \sup D_\rho + v_\rho \,,$$

où R_ρ et D_ρ sont des ensembles définis par :

$$R_\rho = \left\{ r : \mathrm{kl}\left(\mathbf{R}_S(G_\rho) \middle\| r\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right] \right\},$$
$$D_\rho = \left\{ d : \mathrm{kl}\left(\frac{\mathrm{dis}_\rho(S,T)+1}{2} \middle\| \frac{d+1}{2}\right) \leq \frac{1}{m}\left[2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m}}{\delta}\right] \right\}.$$

Démonstration. Le résultat est obtenu en insérant le corollaire 3.1 et le théorème 7.2 dans le théorème 7.4 avec $\delta = \frac{\delta}{2}$. □

Notons que cette borne reste difficile à manipuler en pratique, puisqu'il faut calculer un maximum sur R_ρ et un sur D_ρ.

Toutefois, à des fins algorithmiques, nous allons plutôt chercher à minimiser la borne suivante "à la Catoni", nous permettant de jouer sur le compromis erreur source, ρ-désaccord et KL-divergence.

Théorème 7.7 *Pour tous domaines P_S et P_T sur $X \times Y$ (de marginales selon X respectives D_S et D_T), pour tout ensemble de classifieurs $\mathcal{H}$ de X vers Y, pour toute distribution π sur $\mathcal{H}$, pour*

6. Avec quelques étiquettes cibles, nous pourrions imaginer estimer v_ρ, mais rendrait la borne plus complexe à minimiser. Un autres aspect serait de jouer sur la distribution prior π comme nous le discuterons dans la synthèse.

tous réels $A > 0$, $C > 0$ et pour tout $\delta \in (0, 1]$, avec une probabilité de $1 - \delta$ sur le choix aléatoire d'échantillons de même tailles $S \sim (P_S)^m$, $S_u \sim (D_S)^m$ et $T_u \sim (D_T)^m$, pour toute distribution ρ sur $\mathcal{H}$, on a :

$$\mathbf{R}_{P_T}(G_\rho) \leq A' - 1 + C'\mathbf{R}_S(G_\rho) + A' \operatorname{dis}_\rho(S_u, T_u) + \left(\frac{C'}{C} + \frac{2\,A'}{A}\right) \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{\delta}}{m} + v_\rho,$$

où $C' = \dfrac{C}{1 - e^{-C}}$, *et* $A' = \dfrac{2A}{1 - e^{-2A}}$.

Démonstration. Dans le théorème 7.3, on remplace $\mathbf{R}_S(G_\rho)$ et $\operatorname{dis}_\rho(S_u, T_u)$ par leur majoration, obtenue dans le corollaire 3.3 et le théorème 7.4, avec δ choisi respectivement comme $\frac{\delta}{3}$ et $\frac{2\delta}{3}$ (dans le dernier cas, on utilise $\ln\frac{2}{2\delta/3} = \ln\frac{3}{\delta} < 2\ln\frac{3}{\delta}$). □

Nous voyons apparaître deux paramètres de contrôle de compromis. La section suivante présente un premier algorithme basé sur ce dernier résultat.

7.2 PBDA : adaptation de domaine PAC-Bayésienne spécialisée aux classifieurs linéaires

7.2.1 Formulation générale de l'algorithme

Définissons maintenant $\mathcal{H}$ comme étant un ensemble de classifieurs linéaires de la forme $h(\mathbf{x}) = \operatorname{sign}\langle \mathbf{v}, \mathbf{x}\rangle$ tel que $\mathbf{v} \in \mathbb{R}^d$ est un vecteur de poids.

Nous suivons un processus similaire à la section 3.3 du chapitre 3 dans laquelle nous avons présenté les travaux de [Langford et Shawe-Taylor, 2002, Ambroladze *et al.*, 2006] qui ont spécialisé la théorie PAC-Bayésienne dans le but de borner l'erreur réelle de tout classifieur linéaire identifié par un vecteur de poids $\mathbf{w}$. Nous rappelons que $\pi_\mathbf{0}$ et $\rho_\mathbf{w}$ sont respectivement le prior et le posterior définis comme une gaussienne sphérique de matrice de covariance égale à l'identité centrée respectivement sur les vecteurs $\mathbf{0}$ et $\mathbf{w}$. D'une manière plus formelle, pour tout $h \in \mathcal{H}$ on a :

$$\pi_\mathbf{0}(h) = \left(\frac{1}{\sqrt{2\Pi}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{v}\|^2\right),$$
$$\rho_\mathbf{w}(h) = \left(\frac{1}{\sqrt{2\Pi}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{w}\|^2\right).$$

L'erreur réelle du classifieur de Gibbs $G_{\rho_\mathbf{w}}$ sur un domaine P est alors donnée par :

$$\begin{aligned}\mathbf{R}_{P_S}(G_{\rho_\mathbf{w}}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathop{\mathbf{E}}_{h\sim\rho_\mathbf{w}} \mathbf{I}\left[h(\mathbf{x}) \neq y\right] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \ell_{\mathbf{Erf}}\left(y\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right),\end{aligned}$$

où :

$$\ell_{\mathbf{Erf}}\left(G_{\rho_\mathbf{w}}(\mathbf{x}), y\right) = \frac{1}{2}\left[1 - \mathbf{Erf}\left(\frac{1}{\sqrt{2}}\frac{y\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right)\right],$$

et $\mathbf{Erf}(a) = \frac{2}{\sqrt{\Pi}} \int_0^a \exp\left(-t^2\right) dt$ est la fonction d'erreur de Gauss. Pour rappel, cette situation permet de calculer facilement la KL-divergence entre $\rho_{\mathbf{w}}$ et $\pi_{\mathbf{0}}$:

$$\mathrm{KL}(\rho_{\mathbf{w}} \| \pi_{\mathbf{0}}) = \frac{1}{2}\|\mathbf{w}\|^2.$$

En négligeant la quantité non estimable $v_{\rho_{\mathbf{w}}}$ du théorème 7.7, nous définissons un algorithme PAC-Bayésien pour l'adaptation de domaine inspiré de l'algorithme d'apprentissage de classifieur linéaire PBGD3 [Germain *et al.*, 2009a] présenté en section 3.3 du chapitre 3. Ainsi, étant donnés un échantillon source étiqueté $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ constitué de m exemples *i.i.d.* selon P_S et un échantillon cible non étiqueté $T = \{(\mathbf{x}_i^t)\}_{i=1}^m$ composé de m exemples *i.i.d.* selon D_T, en posant $S_u = \{\mathbf{x}_i^s\}_{i=1}^m$ l'échantillon S privé de ses étiquettes, on veut minimiser :

$$\min_{\mathbf{w}} \; c\, m\, \mathbf{R}_S(G_{\rho_{\mathbf{w}}}) + a\, m\, \mathrm{dis}_{\rho_{\alpha'}}(S_u, T_u) + \mathrm{KL}(\rho_{\mathbf{w}} \| \pi_{\mathbf{0}}), \tag{7.5}$$

où $\mathrm{dis}_{\rho_{\alpha'}}(S_u, T_u) = \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathbf{R}_{S_u}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathbf{R}_{T_u}(h,h') \right|$ est le désaccord empirique entre S_u et T_u spécialisé à une distribution $\rho_{\mathbf{w}}$ sur l'espace $\mathcal{H}$ des classifieurs linéaires considéré. Les réels $a > 0$ et $c > 0$ sont des hyperparamètres de l'algorithme. Notons que les constantes A et C du théorème 7.7 peuvent être retrouvées à partir de n'importe quelle valeur de a et c. Étant donnée la fonction $\ell_{\mathrm{dis}}(x) = 2\,\ell_{\mathbf{Erf}}(x)\,\ell_{\mathbf{Erf}}(-x)$ (illustrée sur la figure 7.1) , pour toute distribution D sur X, on a :

$$\begin{aligned}
\mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathbf{R}_D(h,h') &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}^2}} \mathbf{I}\left[h(\mathbf{x}) = 1\right] \mathbf{I}\left[h'(\mathbf{x}) = -1\right] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathbf{I}\left[h(\mathbf{x}) = 1\right] \mathop{\mathbf{E}}_{h'\sim\rho_{\mathbf{w}}} \mathbf{I}\left[h'(\mathbf{x}) = -1\right] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \ell_{\mathbf{Erf}}\left(\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right) \ell_{\mathbf{Erf}}\left(-\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}\rangle}{\|\mathbf{x}\|}\right).
\end{aligned}$$

Ainsi, trouver la solution optimale de l'équation (7.5) revient à chercher le vecteur $\mathbf{w}$ qui minimise :

$$c \sum_{i=1}^m \ell_{\mathbf{Erf}}\left(y_i^s \frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) + a \left| \sum_{i=1}^m \left[\ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) - \ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^t\rangle}{\|\mathbf{x}_i^t\|}\right)\right]\right| + \frac{\|\mathbf{w}\|^2}{2}. \tag{7.6}$$

L'équation précédente est fortement non convexe. Afin de rendre sa résolution plus facilement contrôlable, nous remplaçons la fonction $\ell_{\mathbf{Erf}}(\cdot)$ par sa relaxation convexe $\ell_{\mathbf{Erf}_{\mathrm{cvx}}}(\cdot)$ (comme pour PBGD3 et illustrée sur la figure 7.1). L'optimisation se réalise ensuite par une descente de gradient. Le gradient de l'équation 7.6 étant :

$$\begin{aligned}
&\mathbf{w} + c \sum_{i=1}^m \ell'_{\mathbf{Erf}_{\mathrm{cvx}}}\left(\frac{\langle y_i^s \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) \frac{y_i^s \mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} \\
&+ s\, a \left(\sum_{i=1}^m \left[\ell'_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^t\rangle}{\|\mathbf{x}_i^t\|}\right) \frac{\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|} - \ell'_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) \frac{\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right]\right),
\end{aligned}$$

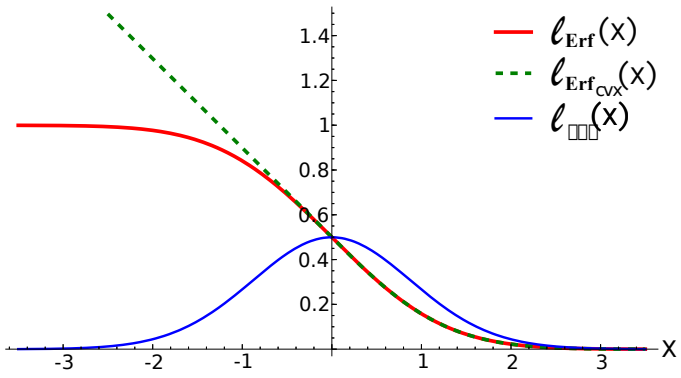


FIGURE 7.1 – *Comportement des fonctions* $\ell_{\mathbf{Erf}}(\cdot)$, $\ell_{\mathbf{Erf}_{\mathrm{cvx}}}(\cdot)$ *et* $\ell_{\mathrm{dis}}(\cdot)$.

où $\ell'_{\mathbf{Erf}_{\mathrm{cvx}}}(x)$, respectivement $\ell'_{\mathrm{dis}}(x)$, est la valeur de la fonction dérivée de $\ell_{\mathbf{Erf}_{\mathrm{cvx}}}(\cdot)$, respectivement $\ell_{\mathrm{dis}}(\cdot)$, évaluée au point x, et :

$$s = \operatorname{sign}\left(\sum_{i=1}^{m}\left[\ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^s\rangle}{\|\mathbf{x}_i^s\|}\right) - \ell_{\mathrm{dis}}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^t\rangle}{\|\mathbf{x}_i^t\|}\right)\right]\right).$$

Cette tâche d'optimisation reste non convexe, puisque la fonction $\ell_{\mathrm{dis}}(\cdot)$ est quasi-concave, cependant, notre étude empirique montre qu'il n'est pas nécessaire de procéder à plusieurs redémarrages pour trouver une solution correcte. Nous appelons cet algorithme d'adaptation de domaine PBDA (pour *PAC-Bayesian Domain Adaptation*).

7.2.2 Utilisation de l'astuce du noyau

L'astuce du noyau présentée en section 1.4.2 du chapitre 1, peut être appliquée à cet algorithme. Cette astuce nous permet de travailler avec le vecteur dual $\boldsymbol{\alpha} \in \mathbb{R}^{2m}$, qui est un classifieur linéaire dans un espace augmenté. Étant donné une fonction noyau $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, on a :

$$h_{\mathbf{w}}(\mathbf{x}) \;=\; \sum_{i=1}^{m} \alpha_i K(\mathbf{x}_i^s, \mathbf{x}) + \sum_{i=1}^{m} \alpha_{i+m} K(\mathbf{x}_i^t, \mathbf{x})\,.$$

On note $\mathbf{K}$ la matrice noyau de dimension $2m \times 2m$ telle que :

$$\forall (i,j) \in \{1,\ldots,2m\}^2, K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$$

où :

$$\mathbf{x}_{\#} = \begin{cases} \mathbf{x}_{\#}^s & \text{si } \# \leq m \\ \mathbf{x}_{\#-m}^t & \text{sinon.} \end{cases}$$

Dans cette situation, la fonction objectif de l'équation (7.6) peut se réécrire en fonction du vecteur $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{2m})^\top$ comme :

$$\frac{1}{2}\sum_{i=1}^{2m}\sum_{j=1}^{2m}\alpha_i\alpha_j\mathbf{K}_{ij} + c\sum_{i=1}^{m}\ell_{\mathbf{Erf}_{\mathrm{cvx}}}\left(y_i^s\frac{\sum_{j=1}^{2m}\alpha_j\mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}}}\right)$$
$$+\, a\left|\sum_{i=1}^{m}\ell_{\mathrm{dis}}\left(\frac{\sum_{j=1}^{2m}\alpha_j\mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}}}\right) - \ell_{\mathrm{dis}}\left(\frac{\sum_{j=1}^{2m}\alpha_j\mathbf{K}_{i+m,j}}{\sqrt{\mathbf{K}_{i+m,i+m}}}\right)\right|.$$

Le gradient de l'équation précédente est donnée par le vecteur $\boldsymbol{\alpha}' = (\alpha'_1, \alpha'_2, \dots \alpha'_{2m})^\top$, avec $\alpha'_\#$ égal à :

$$\sum_{j=1}^{2m} \alpha_i \mathbf{K}_{i,\#} + c \sum_{i=1}^{m} \ell'_{\mathbf{Erf}_{\text{cvx}}} \left(y_i^s \frac{\sum_{j=1}^{2m} \alpha_j \mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}}} \right) \frac{y_i^s \, \mathbf{K}_{i\#}}{\sqrt{\mathbf{K}_{ii}}}$$
$$+ s\, a \left[\sum_{i=1}^{m} \ell'_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j \mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}}} \right) \frac{\mathbf{K}_{i\#}}{\sqrt{\mathbf{K}_{ii}}} - \ell'_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j \mathbf{K}_{(i+m)j}}{\sqrt{\mathbf{K}_{(i+m)(i+m)}}} \right) \frac{\mathbf{K}_{(i+m)\#}}{\sqrt{\mathbf{K}_{(i+m)(i+m)}}} \right],$$

où :

$$s = \text{sign} \left[\sum_{i=1}^{m} \ell_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j \mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}}} \right) - \ell_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j \mathbf{K}_{(i+m)j}}{\sqrt{\mathbf{K}_{(i+m)(i+m)}}} \right) \right].$$

7.3 Expérimentations

7.3.1 Protocole expérimental

Nous avons évalué PBDA sur le jeu de données synthétique des lunes jumelles utilisé dans le chapitre précédent, ainsi que sur un jeu de données d'analyse d'avis populaire en adaptation de domaine. Il a été comparé avec des méthodes non adaptatives : la version de PBGD3 présentée en section 3.3 [Germain *et al.*, 2009a] et un SVM (section 1.4.2, chapitre 1) ; et des méthodes adaptatives : notre DASF du chapitre précédent, DASVM présenté en section 2.3.1[Bruzzone et Marconcini, 2010] , la méthode de co-apprentissage pour l'adaptation de domaine CODA présentée en section 2.3.2. Signalons que dans [Chen *et al.*, 2011a], CODA a montré les meilleurs résultats sur le jeu de données d'analyse d'avis. La fonction objectif de PBDA est minimisée avec une méthode BFGS[7] implémentée dans la librairie python *scipy*[8]. Nous avons fait appel à la bibliothèque SVM-light [Joachims, 1999] pour SVM, DASVM est implémenté avec la bibliothèque LibSVM [Chang et Lin, 2001] et nous avons utilisé l'implémentation[9] proposée par [Chen *et al.*, 2011a] pour CODA. Les paramètres sont sélectionnés à partir d'une grille de recherche via une validation croisée avec 5 sous-ensembles sur l'échantillon source pour PBGD3 et SVM, et via une validation inverse (voir section 2.3.3 du chapitre 2) avec 5 sous-ensembles sur l'échantillon cible (non étiqueté) pour DASF, DASVM, CODA et PBDA.

7.3.2 Problème jouet synthétique

Nous faisons appel au même protocole que dans le chapitre précédent : nous considérons un domaine source et 8 domaines cibles différents en fonction de 8 rotations. Pour chaque domaine, nous générons 300 instances (150 de chaque classe). La capacité en généralisation des algorithmes est évaluée sur un échantillon de test composé de 1 500 exemples cibles. Chacun des problèmes d'adaptation de domaine

7. BFGS est la méthode de Broyden-Fletcher-Goldfarb-Shanno.

8. *scipy* est disponible ici : `http://www.scipy.org/`

9. L'implémentation de CODA est disponible ici : `http://www1.cse.wustl.edu/~mchen/code/coda.tar`.

Angle de rotation	20°	30°	40°	50°	60°	70°	80°	90°
PBGD3	0.088	0.210	0.273	0.399	0.568	0.776	0.804	0.824
SVM	0.104	0.24	0.312	0.4	0.565	0.764	0.808	0.828
DASVM	**0**	0.259	0.284	*0.334*	*0.375*	0.747	0.790	0.820
DASF	0.002	**0.004**	**0.090**	**0.187**	**0.348**	**0.380**	**0.391**	**0.402**
PBDA	0.094	*0.103*	*0.225*	0.412	0.576	*0.626*	*0.604*	*0.687*

TABLE 7.1 – **Problème jouet.** *Taux d'erreur moyens pour les* 8 *angles de rotations*

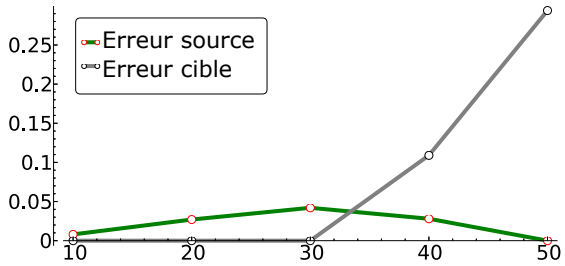
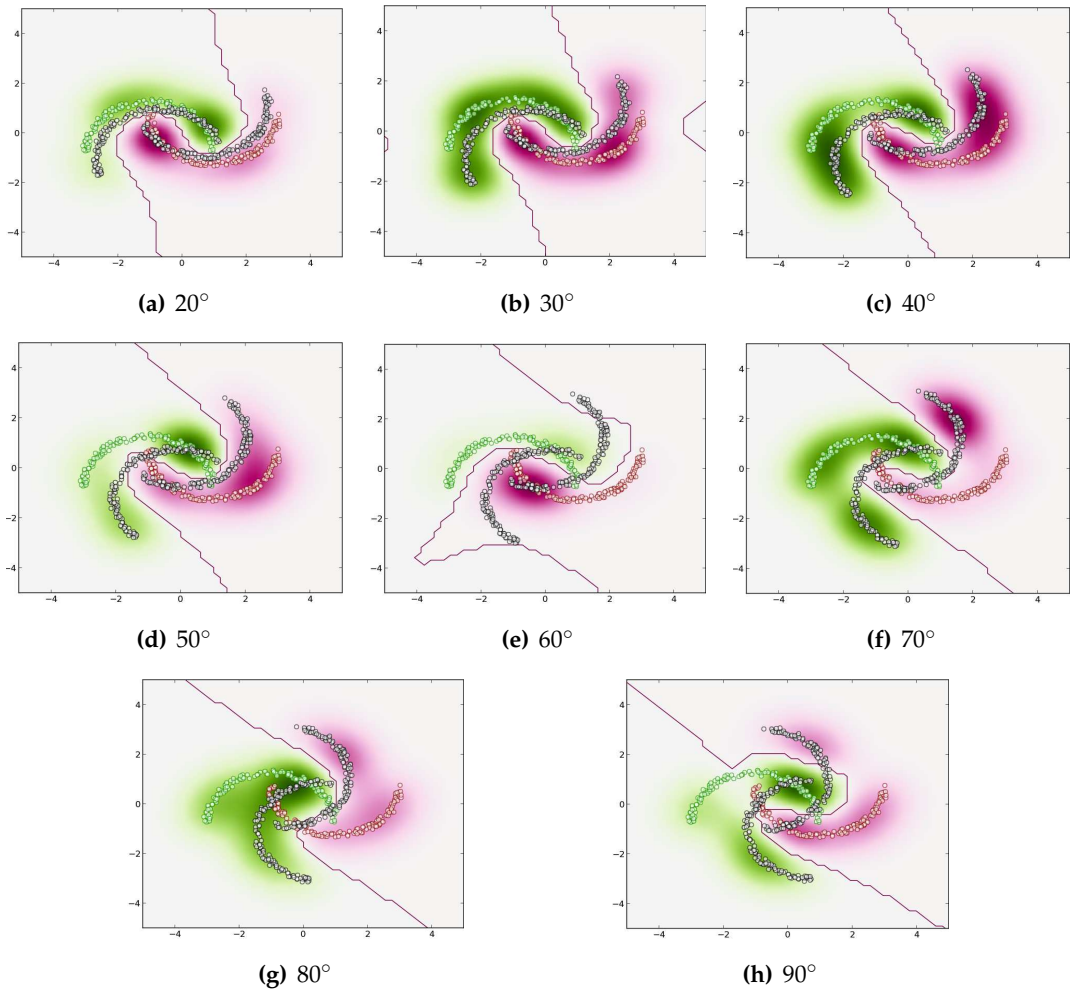


FIGURE 7.2 – **Problème jouet.** *Le compromis erreur source versus erreur cible à paramètres fixés* ($a = c = 1$)*. En abscisse on observe l'angle de rotation et en ordonnée les taux d'erreur.*

est répété 10 fois et nous reportons les résultats moyens dans la table 7.1. Notons que puisque la dimension des données est de 2, CODA, qui décompose les attributs afin d'appliquer un co-apprentissage, ne s'avère pas approprié dans cette expérience. Le noyau utilisé est un noyau Gaussien comme défini dans l'équation (1.12) du chapitre 1. Dans la table 7.1, DASF montre les meilleurs résultats sur ce jeu de données. Cependant, comme nous le verrons dans la section 7.3.3, ses performances se dégradent sur la tâche d'analyse d'avis. En laissant de côté les résultats de DASF, PBDA montre les meilleures performances à l'exception des angles de 20°, 50° et 60°. PBDA est donc un algorithme pertinent pour s'attaquer à une tâche d'adaptation de domaine. Il montre une bonne capacité d'adaptation, en particulier pour les tâches les plus difficiles, probablement car notre divergence de ρ-désaccord entre les distributions marginales $\mathrm{dis}_\rho(D_S, D_T)$ est plus précise et apparaît comme une bonne co-régularisation dans une situation d'adaptation de domaine. Ceci est confirmé par la figures 7.2 et 7.3 : la première met en évidence l'adaptabilité face à la minimisation du risque suggérée par le théorème 7.7, la seconde met en évidence l'évolution de la frontière de décision de l'algorithme pour chacun des angles. En effet, le graphique illustre que PBDA accepte de perdre en performance sur le domaine source pour maintenir sa performance sur le domaine cible (au moins lorsque que les deux domaines ne sont pas trop différents).

7.3.3 Analyse d'avis

Nous considérons le jeu de données appelé *Amazon reviews* [Blitzer *et al.*, 2006] composé d'avis sur quatre types de produits de *Amazon.com*© (*books*, *DVDs*, *electronics*, *kitchen appliances*). À l'origine, les avis s'expriment à l'aide d'étoiles (de 1 à 5) et la



(a) 20° **(b)** 30° **(c)** 40°

(d) 50° **(e)** 60° **(f)** 70°

(g) 80° **(h)** 90°

FIGURE 7.3 – **Problème jouet.** *Illustration de la frontière de décision de PBDA en fixant les hyperparamètres ($a = c = 1$). Le vert et le rose correspondent à l'échantillon source, le gris à l'échantillon cible.*

dimension de la description des données (des uni-grammes et des bi-grammes) est en moyenne de 100 000. Pour simplifier la tâche de classification, nous préférons suivre un protocole similaire à celui proposé par [Chen *et al.*, 2011a]. Les deux classes possibles sont alors : $+1$ pour les produits ayant au moins 4 étoiles, -1 pour ceux ayant au plus 3 étoiles. Un domaine correspond à un type de produit, ce qui implique 12 problèmes d'adaptation de domaine. Par exemple, "*books→DVDs*" correspond à la tâche pour laquelle, *books* est le domaine source et *DVDs* le domaine cible. La dimension des données est réduite de la manière suivante : étant donnée une tâche d'adaptation de domaine, [Chen *et al.*, 2011a] ont uniquement gardé les attributs qui apparaissent au moins 10 fois dans les deux domaines (il reste environ 40 000 attributs), puis ont appliqué une pondération de type tf-idf. Les algorithmes font appel à un noyau linéaire (voir équation (1.11) du chapitre 1) et considèrent 2 000 exemples sources étiquetés et 2 000 cibles non étiquetés. Nous les évaluons sur les ensembles cibles de test proposés par [Chen *et al.*, 2011a] (entre 3 000 et 6 000 exemples), puis nous reportons les résultats sur la table 7.2. Nous faisons les remarques suivantes. Tout d'abord, remarquons que DASF, qui a montré les meilleurs résultats sur le problème jouet, fournit les plus mau-

	B→D	B→E	B→K	D→B	D→E	D→K
PBGD3	*0.174*	0.275	0.236	*0.192*	0.256	0.211
SVM	0.179	0.290	0.251	0.203	0.269	0.232
DASVM	0.193	*0.226*	*0.179*	0.202	*0.186*	0.183
CODA	0.181	0.232	0.215	0.217	0.214	*0.181*
DASF	0.245	0.305	0.230	0.300	0.305	0.247
PBDA	0.183	0.263	0.229	0.197	0.241	0.186

	E→B	E→D	E→K	K→B	K→D	K→E	Moy
PBGD3	0.268	0.245	*0.127*	0.255	0.244	0.235	0.226
SVM	0.287	0.267	0.129	0.267	0.253	0.149	0.231
DASVM	0.305	*0.214*	0.149	0.259	*0.198*	0.157	*0.204*
CODA	0.275	0.239	0.134	*0.247*	0.238	0.153	0.210
DASF	0.330	0.327	0.160	0.340	0.301	0.200	0.275
PBDA	*0.232*	0.221	0.141	*0.247*	0.233	*0.129*	0.208

TABLE 7.2 – **Analyse d'avis.** *Taux d'erreur obtenus sur le domaine cible. B, D, E, K correspondent respectivement au domaine books, DVDs, electronics, kitchen.*

vais résultats : les méthodes non adaptatives sont plus performantes. Ceci s'explique probablement du fait que l'espace de projection induit n'est pas assez expressif au regard de la tâche. Néanmoins, comme espéré, les autres approches adaptatives renvoient les meilleurs résultats en moyennes. Ensuite, PBDA est en moyenne plus performant que CODA, mais moins performant que DASVM. Cependant, PBDA reste compétitif : les résultats ne sont pas significativement différents de ceux obtenus pour CODA et DASVM. De plus, nous avons observé que PBDA est significativement plus rapide que CODA et DASVM : ces deux algorithmes sont basés sur une méthode itérative très coûteuse augmentant le temps d'exécution d'au moins d'un facteur 5 (par rapport à PBDA). En fait, un avantage certain de PBDA est qu'il permet d'optimiser conjointement les termes de notre borne en une seule étape. L'approche PAC-Bayésienne semble donc être pertinente dans le contexte de l'adaptation de domaine, on peut donc imaginer améliorer PBDA en utilisant tous les outils offerts par la théorie PAC-Bayésienne comme discuté ci-dessous.

7.4 Synthèse

Dans ce chapitre, nous avons défini une nouvelle mesure de divergence entre distributions basée sur l'espérance des désaccords sur l'ensemble des votants $\mathcal{H}$. Nous avons démontré des bornes en généralisation justifiant de son estimation empirique simple.

Cette mesure nous a permis de dériver la première analyse PAC-Bayésienne de l'adaptation de domaine. De plus, nous avons proposé un premier algorithme compétitif, plus rapide et fondé théoriquement qui optimise simultanément les termes liés à la divergence et à l'erreur source dans le cas de classifieurs linéaires. Nous sommes persuadés que cette analyse ouvre la porte à de nouvelles méthodes d'adaptation faisant appel aux possibilités offertes par la théorie PAC-Bayésienne, ce qui donne lieu à de nombreuses questions intéressantes.

Comme nous l'avons effectivement vu dans les chapitres 3 et 4, un des intérêts de l'approche PAC-Bayésienne est de pouvoir considérer une connaissance *a priori* sur la performance des classifieurs. Or, ici, nous avons opté pour un prior non informatif (une gaussienne centrée à l'origine de l'espace des classifieurs linéaires). La définition d'un prior pertinent pour l'adaptation de domaine est à étudier, par exemple lorsque quelques étiquettes cibles sont accessibles, ou lorsque différents domaines sources sont disponibles. En effet, un prior dépendant des données étiquetées pourrait nous permettrent de dériver de nouvelles garanties. Une autre direction prometteuse concerne la sélection des hyperparamètres. En effet, l'adaptabilité de notre méthode pourrait être améliorée via une procédure de validation spécifique au cadre PAC-Bayésien. Une idée serait alors de considérer une technique de validation inverse tirant bénéfice des distribution prior et posterior. Enfin une dernière question, serait de mettre en parallèle nos résultats avec ceux de la théorie de *Occam's Hammer* [Blanchard et Fleuret, 2007] qui peut être vue comme une extension de la théorie PAC-Bayésienne à différents types d'objets, et pourrait donc nous permettre de proposer une analyse plus générale.

Conclusion et perspectives

Dans cette thèse, nous avons étudié l'apprentissage de votes de majorité de classifieurs ou de similarités pour des tâches de classification. Nous sommes, dans un premier temps, restés dans le cadre usuel où les données d'apprentissage suivent la même distribution de probabilité que les données de test à classer. Nous avons, ensuite, abordé la situation de l'adaptation de domaine pour laquelle ces deux distributions sont différentes. Nous résumons, ici, un peu plus en détail nos contributions, puis nous nous focalisons sur les perspectives ouvertes par ces travaux.

Notre première contribution a étudié et étendu un algorithme récent qui tire sa source de la théorie PAC-Bayésienne : MinCq. Cet algorithme élégant apprend un vote de majorité pondéré, sur un ensemble de votants à valeurs réelles, en minimisant la C-borne qui majore l'erreur de ce vote [1]. Cette optimisation se réalise à l'aide d'un programme quadratique simple. D'un point de vue théorique, MinCq souffrait, d'une part, de l'impossibilité de considérer un *a priori* sur la distribution des votants et, d'autre part, de l'absence de borne en généralisation valable lorsque les votants dépendent des données d'apprentissage. Nous avons comblé ces lacunes en permettant la prise en considération, durant la phase d'apprentissage, d'une connaissance *a priori* sur la pertinence des votants. De plus, nous avons proposé une preuve de consistance PAC-Bayésienne pour les schémas de compression, c'est-à-dire pour des votants dépendants des données. Nous avons démontré empiriquement l'utilité de notre extension, appelée P-MinCq, pour deux tâches distinctes.

Puisque MinCq est un algorithme limité à la classification binaire, nous avons entrepris, dans notre deuxième contribution, d'étudier la classification multiclasse sous un angle PAC-Bayésien. Nous avons, dans un premier temps, démontré la première borne en généralisation PAC-Bayésienne portant sur le moyennage des matrices de confusion des classifieurs à combiner. Dans un second temps, avec l'objectif de proposer une approche de type P-MinCq/MinCq pour apprendre un vote de majorité multiclasse, nous avons généralisé la C-borne à ce contexte. Cette première série de résultats théoriques n'a pas permis de concevoir un algorithme capable d'obtenir des performances similaires à l'état de l'art. Néanmoins, nous pensons que ces travaux ouvrent des pistes intéressantes pour la classification multiclasse.

1. Nous rappelons que la C-borne met en jeu le premier et le second moment de la marge du vote de majorité.

Pour notre troisième contribution, nous nous sommes penchés sur la problématique de l'adaptation de domaine pour la classification binaire. Cette contribution tire à la fois bénéfice de la théorie de l'apprentissage avec une fonction de similarité (ϵ,γ,τ)-bonne et de la théorie classique de l'adaptation de domaine. (SS)DASF, l'algorithme que nous avons proposé, apprend un vote de majorité sur un ensemble de similarités en repondérant itérativement les similarités. Ce schéma de repondération est formulé comme une co-régularisation du problème d'optimisation à résoudre, dont le but est de rapprocher les domaines tout en gardant de bonnes garanties sur les données d'apprentissage. Le vote final prend alors la forme d'un classifieur linéaire dans un espace de projection explicitement défini par les similarités. Notre approche permet de s'attaquer aux situations pour lesquelles aucune information supervisée sur les étiquettes cibles n'est disponible, mais aussi celles où quelques étiquettes cibles sont accessibles. Cet algorithme a, empiriquement, démontré de bonnes capacités d'adaptation.

Enfin, pour répondre aux deux problématiques principales soulevées dans ce mémoire, nous avons analysé — pour la première fois — l'adaptation de domaine avec une approche PAC-Bayésienne. Concrètement, alors que la théorie classique calcule la divergence entre les domaines dans le pire des cas [2], notre dernière contribution se fonde sur une nouvelle mesure, plus précise et plus simple à estimer. Elle s'avère totalement adaptée à la théorie PAC-Bayésienne et, donc, à l'apprentissage de vote de majorité. En effet, elle nous a permis de dériver des bornes en généralisation PAC-Bayésienne pour l'adaptation de domaine. Une de ces bornes a l'avantage de pouvoir être optimisée directement pour un ensemble de classifieurs et suggère une nouvelle approche d'adaptation : on peut combiner des classifieurs — ou d'autres fonctions — en minimisant simplement un compromis entre la complexité du vote de majorité mesuré par la KL-divergence, son risque empirique et sa capacité à discerner les domaines. Nous l'avons spécialisée aux cas des classifieurs linéaires pour proposer un premier algorithme d'adaptation PAC-Bayésien qui s'est montré des plus prometteurs.

En restant dans le champ d'étude du sujet de cette thèse, l'apprentissage d'un vote de majorité pondéré performant et profitant de la diversité des fonctions à combiner doit très probablement tirer bénéfice de la C-borne. C'est pourquoi, une des premières perspectives qui s'offre à nous est la généralisation de la C-borne et de l'algorithme P-MinCq/MinCq à la classification multiclasse mais aussi à l'adaptation de domaine. En effet, pour cette dernière, la C-borne, qui considère les deux premiers moments de la marge du vote, pourrait nous aider à utiliser à la fois une information sur la marge sur les données non étiquetées, mais aussi sur le désaccord entre classifieurs : ce sont deux éléments qui semblent être d'une importance cruciale en adaptation de domaine.

En outre, nous avons uniquement étudié des problèmes pour lesquels on ne considère qu'un seul domaine source. S'intéresser à l'adaptation de domaine multi-source [Crammer *et al.*, 2008] est une perspective importante. D'une part, avec encore une

2. Nous rappelons que la $\mathcal{H}$-divergence s'exprime avec un *supremum* et est difficile à optimiser simultanément avec l'erreur source.

fois l'intuition selon laquelle l'information du second ordre et, donc, le désaccord entre classifieurs est nécessaire à un bon processus d'adaptation, l'approche P-MinCq/MinCq se doit d'être étudiée dans une telle situation. D'autre part, nous pourrions étendre directement nos travaux à l'adaptation multi-source. Une première solution serait de partir de l'étude théorique basée sur la $\mathcal{H}$-divergence et proposée par [Blitzer *et al.*, 2007, Ben-David *et al.*, 2010]. Une autre direction serait d'étudier plus en détail les divergences utilisées par [Mansour *et al.*, 2009b, C. Zhang, 2012] et explicitement développées dans un contexte multi-source. De plus, nous pensons qu'il est possible d'étendre aisément notre mesure de ρ-désaccord à plusieurs domaines. Une telle extension permettrait d'obtenir des bornes en généralisation proches des résultats obtenus dans le dernier chapitre.

De surcroît, il est possible que le domaine cible change ou évolue au cours du temps. Ce problème, connu sous le nom de *distribution drift* ou de *concept drift* [Žliobaitė, 2010], est donc à étudier. Comme l'ont proposé [Mohri et Medina, 2012] pour la *discrepancy*, une première stratégie serait d'exploiter notre ρ-désaccord pour en définir une généralisation à cette situation. Une seconde piste serait de tirer avantage de la théorie PAC-Bayésienne et, en particulier, de la possibilité de prendre en considération un *a priori* sur la distribution des votants. En effet, nous pourrions apprendre le nouveau posterior en considérant, comme prior, le posterior de l'étape précédente.

De plus, cette notion de prior dans l'analyse PAC-Bayésienne révèle un avantage certain. En effet, nous savons que plus cet *a priori* est pertinent, plus la complexité mesurée par la KL-divergence entre le prior et le posterior sera faible et plus les bornes en généralisation seront précises. Citons, par exemple, les bornes obtenues pour les SVM par [Langford et Shawe-Taylor, 2002, Lever *et al.*, 2010, Lever *et al.*, 2013]. Définir un *a priori* spécifique à une tâche donnée est donc une des perspectives sous-jacentes à nos contributions. Par exemple, nous pensons que la combinaison de classifieurs appris à partir de différentes descriptions doit pouvoir être améliorée en ayant défini, en amont de la phase d'apprentissage, un *a priori* pertinent. De la même manière, nous pouvons imaginer que les bornes PAC-Bayésienne pour l'adaptation de domaine peuvent être améliorées à l'aide d'un prior adéquat, mais aussi en adaptant la définition du ρ-désaccord pour tenir compte de cette information *a priori*. Un autre point concerne la validation des hyperparamètres en adaptation de domaine. En effet, nous nous demandons s'il ne serait pas possible de développer une procédure s'aidant de la distribution prior.

Dans les expérimentations menées dans le chapitre 6 pour évaluer (SS)DASF, nous avons vu que l'utilisation de fonctions de similarités ni semi-définies positives, ni symétriques pouvait aider à la résolution d'un problème d'adaptation de domaine. Dans la section 6.4.1, nous avons uniquement fait appel à une simple heuristique pour "adapter" une telle similarité. Cependant, ces résultats prometteurs ouvrent la perspective de l'apprentissage d'une fonction de similarité (ϵ,γ,τ)-bonne étant donnée une tâche d'adaptation. Une des directions possibles est l'investigation d'approches

d'apprentissage de métrique. À ce jour, peu de méthodes existent avec une perspective d'adaptation de domaine [Zhang et Yeung, 2010, Geng *et al.*, 2011, Cao *et al.*, 2011, Kulis *et al.*, 2011]. Elles se focalisent, de surcroît, quasiment toutes sur des fonctions semi-définies positives. Une première solution serait de s'aider des travaux existants en apprentissage de fonctions de similarités non symétriques [Bellet *et al.*, 2011]. En outre, l'approche PAC-Bayésienne pourrait elle aussi apporter un intérêt certain, à la fois pratique et théorique. D'une part, l'apprentissage d'un prior peut être considéré comme l'apprentissage d'une pondération *a priori* des objets à combiner. Dans ce cas, la métrique est définie comme une combinaison pondérée. D'autre part, elle pourrait nous permettre de proposer des garanties PAC-Bayésiennes pour l'apprentissage d'une telle métrique.

Nous avons vu, dans ce mémoire, que les approches d'adaptation de domaine sont toutes basées sur des hypothèses particulières liées aux relations qu'entretiennent les domaines. Ces hypothèses peuvent s'exprimer par exemple soit explicitement par le *covariate-shift*, soit implicitement notamment lorsque les termes associés aux étiquetages dans les bornes en généralisation sont supposés faibles. Une problématique difficile, mais intéressante, serait donc de concevoir d'un cadre adaptatif générique paramétrable en fonction de la tâche. Puisqu'il est crucial de faire appel à une divergence informative et pertinente entre les domaines, l'obtention d'un tel résultat passe très probablement par la définition d'une divergence générale. L'aboutissement en serait la preuve d'une borne en généralisation vérifiable dans la grande majorité des situations et aidant au développement de nouvelles procédures d'adaptation.

D'un point de vue plus général, que ce soit la question de la combinaison de classifieurs dans un but de créer un modèle plus robuste et plus performant, ou celle de la capacité d'adaptation d'un classifieur, elles soulèvent toutes les deux la problématique suivante. Lorsque l'on désire concevoir un système efficace, robuste et capable de s'adapter aux données et au cours du temps, nous devons nous interroger sur le meilleur moyen pour tirer bénéfice de différentes tâches et/ou d'anciennes tâches. Le terme *long-life learning* est parfois utilisé dans ce contexte. Son principe est illustré sur la figure Ccl.1. De notre point de vue, il faut donc être capable d'apprendre des modèles à la fois adaptatifs et multi-tâches. Comme l'est l'adaptation de domaine, l'apprentissage multi-tâche [Evgeniou *et al.*, 2006] est une sous-problématique de l'apprentissage par transfert. L'objectif est d'apprendre en même temps, souvent à l'aide d'une co-régularisation, plusieurs tâches reliées entre elles en faisant appel à une représentation commune. Ceci permet en général de produire des modèles plus performants. L'analyse PAC-Bayésienne de l'adaptation de domaine pourrait nous aider à définir un nouveau cadre théorique pour cette problématique de *long-life learning*. De plus, l'évolution au cours du temps d'un modèle suggère naturellement de se poser la question de l'apprentissage en ligne [Cesa-Bianchi *et al.*, 2004], ou de l'apprentissage actif [Settles, 2010]. Dans les deux cas, nous devons facilement pouvoir adapter des bornes en généralisation PAC-Bayésiennes existantes (voir par exemple [Seldin *et al.*, 2011]).

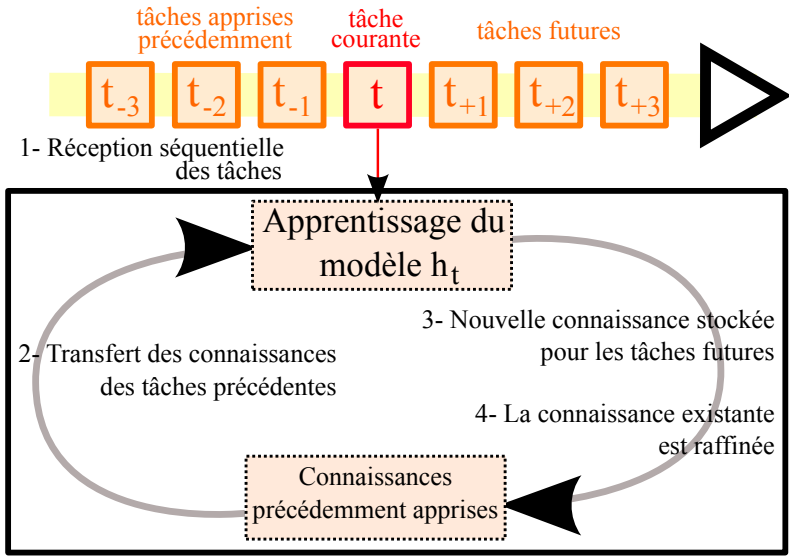


FIGURE Ccl.1 – *Principe du long-life learning.*

Une autre piste serait d'exploiter la théorie des noyaux à valeurs opérateurs (ou noyaux multi-tâche) [Senkene et Tempel'man, 1973, Micchelli et Pontil, 2005] qui permettent de travailler avec des données fonctionnelles et donc avec des représentations ou des sorties plus complexes. Par exemple, nous pourrions traiter la matrice de confusion PAC-Bayésienne via cette théorie. Enfin, quelle que soit la piste envisagée ou l'algorithme utilisé, la question de la validation des paramètres est importante. Une première direction à étudier pour y répondre serait d'expérimenter les travaux proposés par [Bardenet *et al.*, 2013], qui permettent d'incorporer une connaissance *a priori* de tâches précédemment réalisées mais similaires.

Finalement, les travaux présentés dans ce mémoire permettent également de faire le lien avec des problématiques actuelles importantes en apprentissage. Un premier aspect concerne la prédiction structurée lorsque l'on veut prédire des sorties multiples au même moment. Nous pensons que cette problématique peut facilement être reliée au multiclasse, au multi-tâche, voire à la multimodalité. Comme nous l'avons suggéré dans la synthèse du chapitre 5, l'existence d'inégalités de concentration sur les matrices aléatoires qui ne prennent pas en compte la dimension [Hsu *et al.*, 2012], mais aussi de bornes PAC-Bayésiennes spécifiques à ce contexte [Giguère *et al.*, 2013], peuvent aider à généraliser nos résultats multiclasses à la prédiction structurée. Un deuxième aspect est le *deep learning* dont l'un des travaux de référence en adaptation de domaine est [Glorot *et al.*, 2011]. En fait, les méthodes de *deep learning* sont basées sur l'apprentissage de représentation. Or, comme nous l'avons vu, une donnée peut être décrite selon différentes représentations et certaines descriptions sont plus pertinentes pour certaines tâches et/ou certaines données. Il faut donc être capable de définir de bonnes représentations ou de les "apprendre" correctement. Ainsi, tirer avantage de ce cadre pour définir de nouvelles méthodes d'adaptation, de combinaison de votants, ou même de *long-life learning* est une perspective des plus promet-

teuses. Un dernier aspect serait de se focaliser sur l'apprentissage par renforcement [Kaelbling *et al.*, 1996, Sutton et Barto, 1998] dont le contexte est différent. En effet, étant donnée une tâche, son objectif est de trouver des actions pertinentes dans le but de maximiser une récompense. Ce domaine étudie le compromis entre exploration et exploitation. Au lieu de considérer des observations, l'apprenant reçoit une récompense en fonction de son action. Toujours dans une perspective de *long-life learning*, une piste intéressante serait alors de faire appel aux travaux PAC-Bayésiens [Seldin *et al.*, 2011] dans ce domaine pour dériver de nouveaux résultats en adaptation de domaine, mais aussi en multi-tâche et multi-source.

Annexes

Quelques outils

A

Théorème A.1 (Inégalité de Hoeffding) *Soit une séquence finie $\{Z_i\}_{1\leq i\leq n}$ de variables aléatoires réelles indépendantes telles que pour deux séquences $\{a_i\}_{1\leq i\leq n}$ et $\{b_i\}_{1\leq i\leq n}$ de nombres réels avec $a_i < b_i$, on a :*

$$\forall i \in \{1,\ldots,n\},\ \mathbf{Pr}\ (a_i \leq Z_i \leq b_i) = 1.$$

On pose $S_n = Z_1 + \cdots + Z_n$. Alors pour tout $\epsilon > 0$, on a :

$$\mathbf{Pr}\ \left(\left|S_n - \mathbf{E}\ [S_n]\right| \geq \epsilon\right) \leq 2\exp\ \left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Théorème A.2 (Inégalité de Cantelli-Chebitchev) *Soit Z une variable aléatoire.*

$$\forall a \geq 0,\ \mathbf{Pr}\ \left(Z \leq \mathbf{E}\ [Z] - a\right) \leq \frac{\mathbf{Var}\ Z}{\mathbf{Var}\ Z + a^2}.$$

Théorème A.3 (Inégalité de Hölder) *Soient $\mathbf{u}$ et $\mathbf{v}$ deux vecteurs de réels, alors :*

$$\|\mathbf{uv}\|_1 \leq \|\mathbf{u}\|_p \|\mathbf{v}\|_q,$$

avec $1 \leq p,\ q \leq \infty$ et $1/p + 1/q = 1$.

Théorème A.4 (Inégalité de Markov) *Soit Z une variable aléatoire et $\epsilon \geq 0$, alors :*

$$\mathbf{Pr}\ (|Z| \geq \epsilon) \leq \frac{\mathbf{E}\ (|Z|)}{\epsilon}.$$

Théorème A.5 (Inégalité de Jensen) *Soit Z une variable aléatoire réelle et intégrable et $g(\cdot)$ une fonction convexe, alors :*

$$g(\mathbf{E}\ [Z]) \leq \mathbf{E}\ [g(Z)].$$

Lemme A.1 (Issue des inégalités (1) et (2) de [Maurer, 2004]) *Soit $m \geq 8$ et $Z = (Z_1,\ldots,Z_n)$ un vecteur de variables aléatoires i.i.d. telles que $0 \leq Z_i \leq 1$, alors :*

$$\sqrt{n} \leq \mathbf{E}\ \exp\left[n\,\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n Z_i \middle\| \mathbf{E}\ [Z_i]\right)\right] \leq 2\sqrt{n},$$

où $\mathrm{kl}(a\|b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$.

Annexe du chapitre 3

B

B.1 Preuve du théorème PAC-Bayes 3.2

Considérons la variable aléatoire non-négative : $\mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}$. On applique alors l'inégalité de Markov (théorème A.4, annexe A), pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$, on a :

$$\mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))} \leq \frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}.$$

En appliquant la fonction logarithme de chaque côté de cette inégalité, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$ et pour toute distribution posterior ρ sur $\mathcal{H}$, on a :

$$\ln\left[\mathop{\mathbf{E}}_{h\sim\rho} \frac{\pi(h)}{\rho(h)} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}\right] \leq \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}\right].$$

Puisque la fonction $\ln(\cdot)$ est une fonction concave, nous pouvons appliquer l'inégalité de Jensen (théorème A.5, annexe A). Ainsi, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$ et pour toute distribution posterior ρ sur $\mathcal{H}$, on a :

$$\mathop{\mathbf{E}}_{h\sim\rho} \ln\left[\frac{\pi(h)}{\rho(h)} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}\right] \leq \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}\right].$$

D'après la définition de la KL-divergence (équation 3.2), on a :

$$\mathop{\mathbf{E}}_{h\sim\rho} \ln\left[\frac{\pi(h)}{\rho(h)}\right] = -\operatorname{KL}(\rho\|\pi).$$

Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$ et pour toute distribution posterior ρ sur $\mathcal{H}$, on a donc :

$$-\operatorname{KL}(\rho\|\pi) + \mathop{\mathbf{E}}_{h\sim\rho} m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h)) \leq \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}\right].$$

Puisque $\mathcal{D}(\cdot,\cdot)$ est une fonction convexe, l'inégalité de Jensen implique :

$$\mathop{\mathbf{E}}_{h\sim\rho} m\mathcal{D}\left(\mathbf{R}_S(h),\mathbf{R}_P(h)\right) \geq m\mathcal{D}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_S(h), \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_P(h)\right).$$

D'après la définition de l'erreur du classifieur de Gibbs, on a :

$$m\mathcal{D}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_S(h), \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{R}_P(h)\right) = m\mathcal{D}\left(\mathbf{R}_S(G_\rho),\mathbf{R}_P(G_\rho)\right).$$

Finalement, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$ et pour toute distribution posterior ρ sur $\mathcal{H}$, on obtient le résultat du théorème :

$$\mathcal{D}\left(\mathbf{R}_S(G_\rho), \mathbf{R}_P(G_\rho)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))}\right)\right].$$

B.2 Preuve du corollaire 3.3

Soit $\mathcal{F}(\cdot)$ une fonction convexe. Considérons $\mathcal{D}(a,b) = \mathcal{F}(b) - Ca$. On a :

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))} &= \mathop{\mathbf{E}}_{h\sim\pi} \mathop{\mathbf{E}}_{S\sim(P)^m} \mathrm{e}^{m\mathcal{F}(\mathbf{R}_P(h)) - Cm\mathbf{R}_S(h)} \\
&= \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{F}(\mathbf{R}_P(h))} \sum_{k=0}^{m} \mathop{\mathbf{Pr}}_{S\sim(P)^m}\left(\mathbf{R}_S(h) = \frac{k}{m}\right) \mathrm{e}^{-Ck} \\
&= \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{F}(\mathbf{R}_P(h))} \sum_{k=0}^{m} \mathbf{R}_P(h)^k \left(1-\mathbf{R}_P(h)\right)^{m-k} \mathrm{e}^{-Ck} \\
&= \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{F}(\mathbf{R}_P(h))} \left(\mathbf{R}_P(h)\mathrm{e}^{-C} + \left(1-\mathbf{R}_P(h)\right)\right)^m.
\end{aligned}$$

Avec $\mathcal{F}(b) = \ln \dfrac{1}{1-b\left[1-\exp\left(-C\right)\right]}$, on obtient le résultat du corollaire :

$$\mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{h\sim\pi} \mathrm{e}^{m\mathcal{D}(\mathbf{R}_S(h),\mathbf{R}_P(h))} = 1.$$

B.3 Preuve de la C-borne, théorèmes 3.1 et 3.3

Le résultat s'obtient en appliquant l'inégalité de Cantelli-Chebitchev (théorème A.2, annexe A). On remplace Z par la variable aléatoire $\mathcal{M}^\rho(\mathbf{x},y)$ et a par $\mathcal{M}_P^\rho$. D'après la définition 3.2, on a :

$$\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x},y) = \mathcal{M}_P^{\rho^2} - \left(\mathcal{M}_P^\rho\right)^2.$$

Alors :

$$\begin{aligned}
\mathbf{R}_P(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left(\mathcal{M}^\rho(\mathbf{x},y) \leq 0\right) \\
&\leq \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x},y)}{\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x},y) + \left(\mathcal{M}_P^\rho\right)^2} \\
&= \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x},y)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\, \left(\mathcal{M}^\rho(\mathbf{x},y)\right)^2} \\
&= 1 - \frac{\left(\mathcal{M}_P^\rho\right)^2}{\mathcal{M}_P^{\rho^2}}.
\end{aligned}$$

Annexe du chapitre 4

C

C.1 Preuve de la proposition 4.1

Soit ρ une distribution sur $\mathcal{H}$. Soit M tel que :

$$M = \max_{j' \in \{1,\dots,n\}} \frac{1}{\pi_{j'}} \left| \rho(h_{j'+n}) - \rho(h_{j'}) \right|.$$

Soit la distribution ρ' définie par :

$$\rho'(h_j) = \frac{\pi_j}{2} + \frac{\rho(h_j) - \rho(h_{j+n})}{2M},$$

où par convention :

$$(j+n)+n = j\,, \quad \text{et} \quad \pi(h_{j+n}) = \pi(h_j) = \frac{\pi_j}{2}.$$

Montrons tout d'abord que ρ' est en fait une distribution $\boldsymbol{\pi}$-alignée sur $\mathcal{H}$ auto-complémenté, telle que :

$$\forall j \in \{1,\dots,n\},\ \rho'(h_j) \leq \pi_j \text{ et } \rho'(h_j) + \rho'(h_{j+n}) = \pi_j.$$

On a :

$$\begin{aligned}
\rho'(h_j) \leq \pi_j \quad &\iff \quad \frac{\pi(h_j)}{2} + \frac{\rho(h_j) - \rho(h_{j+n})}{2M} \leq \pi_j \\
&\iff \quad \frac{\pi_j}{2} + \frac{\rho(h_j) - \rho(h_{j+n})}{2M} \leq \pi_j \\
&\iff \quad \frac{\rho(h_j) - \rho(h_{j+n})}{M} \leq \pi_j \\
&\iff \quad \frac{1}{\pi_j}\left[\rho(h_j) - \rho(h_{j+n})\right] \leq \max_{j' \in \{1,\dots,n\}} \frac{1}{\pi_{j'}} \left| \rho(h_{j'+n}) - \rho(h_{j'}) \right|,
\end{aligned}$$

De plus :

$$\begin{aligned}
\rho'(h_j) + \rho'(h_{j+n}) &= \frac{\pi(h_j)}{2} + \frac{\rho(h_j) - \rho(h_{j+n})}{2M} + \frac{\pi(h_{j+n})}{2} + \frac{\rho(h_{j+n}) - \rho(h_j)}{2M} \\
&\quad \frac{\pi_j}{2} + \frac{\rho(h_j) - \rho(h_{j+n})}{2M} + \frac{\pi_j}{2} + \frac{\rho(h_{j+n}) - \rho(h_j)}{2M} \\
&= \pi_j + \frac{\rho(h_j) - \rho(h_{j+n}) + \rho(h_{j+n}) - \rho(h_j)}{2M} \\
&= \pi_j.
\end{aligned}$$

Montrons maintenant que ρ' ne restreint pas l'ensemble des votes de majorité possibles :

$$\begin{aligned}
\mathop{\mathbf{E}}_{h\sim\rho'} h(\mathbf{x}) &= \sum_{j=1}^{2n} \rho'(h_j) h_j(\mathbf{x}) \\
&= \sum_{j=1}^{n} \left[\rho'(h_j) - \rho'(h_{j+n})\right] h_j(\mathbf{x}) \\
&= \frac{1}{M} \sum_{j=1}^{n} \left[\rho(h_j) - \rho(h_{j+n})\right] h_j(\mathbf{x}) \\
&= \frac{1}{M} \sum_{j=1}^{2n} \rho(h_j) h_j(\mathbf{x}) \\
&= \frac{1}{M} \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}).
\end{aligned}$$

Ainsi, nous avons : $\forall \mathbf{x} \in X$, $B_{\rho'}(\mathbf{x}) = B_{\rho}(\mathbf{x})$. Puisque la constante $\frac{1}{M}$ est présente dans le premier et le second moment $\mathcal{M}_P^{\rho'}$ et $\mathcal{M}_P^{\rho'}$, elle disparaît dans la C-borne. Donc, $C_{\rho'}^P = C_{\rho}^P$ quelque soit la distribution P sur $X \times Y$.

C.2 Preuve du théorème 4.1

Preuve de l'équation (4.2)**.**

Soit S un ensemble d'apprentissage de taille m. Posons $\mathcal{H}^S$ une famille de votants auto-complémentée. Rappelons que π et ρ sont deux distribution sur $\mathcal{H}^S$. Une distribution de probabilité est $\boldsymbol{\pi}$-alignée sur $\mathcal{H}^S$ si pour tout $(\mathbf{i}, \sigma) \in \mathbf{I}_m \times \Omega_{S_\mathbf{i}}$ on a :

$$\begin{aligned}
\rho\left(h_S^{(\sigma,+)}\right) + \rho\left(-h_S^{(\sigma,+)}\right) &= \rho\left(h_S^{(\sigma,+)}\right) + \rho\left(h_S^{(\sigma,-)}\right) \\
&= \pi\left(h_S^{(\sigma,+)}\right) + \pi\left(h_S^{(\sigma,-)}\right) \\
&= \pi\left(h_S^{(\sigma,+)}\right) + \pi\left(-h_S^{(\sigma,+)}\right).
\end{aligned}$$

En posant :

$$\mathcal{M}_P^{h_S^{(\sigma,+)}} = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} y\, h_S^{(\sigma,+)}(\mathbf{x}), \quad \text{et} \quad \mathcal{M}_S^{h_S^{(\sigma,+)}} = \frac{1}{m} \sum_{i=1}^{m} y_i\, h_S^{(\sigma,+)}(\mathbf{x}_i),$$

on a :

$$\mathcal{M}_P^{h_S^{(\sigma,+)}} = -\mathcal{M}_P^{h_S^{(\sigma,-)}},$$

et :

$$\begin{aligned}
\left(\mathcal{M}_S^{h_{S_\mathbf{i}}^{(\sigma,+)}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^{(\sigma,+)}}\right)^2 &= \left(-\mathcal{M}_S^{h_{S_\mathbf{i}}^{(\sigma,-)}} - \left(-\mathcal{M}_P^{h_{S_\mathbf{i}}^{(\sigma,-)}}\right)\right)^2 \\
&= \left(\mathcal{M}_S^{h_{S_\mathbf{i}}^{(\sigma,-)}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^{(\sigma,-)}}\right)^2.
\end{aligned}$$

Similairement à [McAllester, 2003, Laviolette *et al.*, 2011b], nous considérons la transformée de Laplace suivante :

$$X_\pi = \mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{i}} \sim \pi} \exp\left[\frac{m - |\mathbf{i}|}{2B^2}\left(\mathcal{M}_S^{h^\omega_{S_\mathbf{i}}} - \mathcal{M}_P^{h^\omega_{S_\mathbf{i}}}\right)^2\right],$$

où B borne la valeur des votants. Notons que $f(a,b) = \frac{1}{2B^2}(a-b)^2$ est une fonction convexe car sa matrice hessienne est semi-définie postive. Afin de simplifier la lecture de la preuve nous posons :

$$m_\mathbf{i} = \frac{m - |\mathbf{i}|}{2B^2}.$$

Pour toute distribution $\boldsymbol{\pi}$-alignée ρ sur $\mathcal{H}^S$, on a :

$$\begin{aligned}
2X_\pi =& \mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{i}} \sim \pi} \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^\omega_{S_\mathbf{i}}} - \mathcal{M}_P^{h^\omega_{S_\mathbf{i}}}\right)^2\right] \\
=& \int_{h^{(\sigma,+)}_{S_\mathbf{i}} \in \mathcal{H}^S} \pi\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right) \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}}\right)^2\right] dh^{(\sigma,+)}_{S_\mathbf{i}} \\
&+ \int_{h^{(\sigma,-)}_{S_\mathbf{i}} \in \mathcal{H}^S} \pi\left(h^{(\sigma,-)}_{S_\mathbf{i}}\right) \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^{(\sigma,-)}_{S_\mathbf{i}}} - \mathcal{M}_P^{h^{(\sigma,-)}_{S_\mathbf{i}}}\right)^2\right] dh^{(\sigma,-)}_{S_\mathbf{i}} \\
=& \int_{h^{(\sigma,+)}_{S_\mathbf{i}} \in \mathcal{H}^S} \left[\pi\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right) + \pi\left(-h^{(\sigma,+)}_{S_\mathbf{i}}\right)\right] \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}}\right)^2\right] dh^{(\sigma,+)}_{S_\mathbf{i}} \\
=& \int_{h^{(\sigma,+)}_{S_\mathbf{i}} \in \mathcal{H}^S} \left[\rho\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right) + \rho\left(-h^{(\sigma,+)}_{S_\mathbf{i}}\right)\right] \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}}\right)^2\right] dh^{(\sigma,+)}_{S_\mathbf{i}} \\
=& \int_{h^{(\sigma,+)}_{S_\mathbf{i}} \in \mathcal{H}^S} \rho\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right) \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}}\right)^2\right] dh^{(\sigma,+)}_{S_\mathbf{i}} \\
&+ \int_{h^{(\sigma,-)}_{S_\mathbf{i}} \in \mathcal{H}^S} \rho\left(h^{(\sigma,-)}_{S_\mathbf{i}}\right) \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^{(\sigma,-)}_{S_\mathbf{i}}} - \mathcal{M}_P^{h^{(\sigma,-)}_{S_\mathbf{i}}}\right)^2\right] dh^{(\sigma,-)}_{S_\mathbf{i}} \\
=& 2 \mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{i}} \sim \rho} \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^\omega_{S_\mathbf{i}}} - \mathcal{M}_P^{h^\omega_{S_\mathbf{i}}}\right)^2\right] \\
=& 2X_\rho.
\end{aligned}$$

En appliquant l'inégalité de Markov (théorème A.4, annexe A), pour tout $\delta \in (0,1]$ on obtient :

$$\mathop{\mathbf{Pr}}_{S \sim (P)^m}\left(X_\pi \leq \frac{1}{\delta} \mathop{\mathbf{E}}_{S \sim (P)^m} X_\pi\right) \geq 1 - \delta.$$

On applique le logarithme aux deux côtés de l'inégalité. Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$, pour toute distribution $\boldsymbol{\pi}$-alignée sur $\mathcal{H}^S$, on a :

$$\ln\left(\mathop{\mathbf{E}}_{h^\omega_{S_\mathbf{i}} \sim \rho} \exp\left[m_\mathbf{i}\left(\mathcal{M}_S^{h^\omega_{S_\mathbf{i}}} - \mathcal{M}_P^{h^\omega_{S_\mathbf{i}}}\right)^2\right]\right) \leq \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S \sim (P)^m} X_\pi\right).$$

La fonction $\ln(\cdot)$ est une fonction concave, on peut donc appliquer l'inégalité de Jensen (théorème A.5, annexe A) sur le terme de gauche de l'inégalité précédente :

$$\ln\left(\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{i}}}\sim\rho}\exp\left[m_{\mathbf{i}}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}}}\right)^2\right]\right)\geq\underbrace{\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{i}}}\sim\rho}m_{\mathbf{i}}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}}}\right)^2}_{(\mathbf{g})}.$$

Rappelons que $|\mathbf{i}_{\max}|$ est la taille maximale de la séquence de compression. Alors, en appliquant l'inégalité de Jensen sur $(\mathbf{g})$ avec la fonction convexe $(m-|\mathbf{i}_{\max}|)f(a,b)=\frac{m-|\mathbf{i}_{\max}|}{2B^2}(a-b)^2\leq m_{\mathbf{i}}(a-b)^2$, on obtient :

$$\begin{aligned}(\mathbf{g})=\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{i}}}\sim\rho}m_{\mathbf{i}}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}}}\right)^2&=\frac{m}{2B^2}\left[\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{i}}}\sim\rho}-|\mathbf{i}|\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}}}\right)^2\right]\\&\geq\frac{m-|\mathbf{i}_{\max}|}{2B^2}\left[\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{i}}}\sim\rho}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}}}\right)^2\right]\\&\geq\frac{m-|\mathbf{i}_{\max}|}{2B^2}\left(\mathcal{M}_S^{\rho}-\mathcal{M}_P^{\rho}\right)^2.\end{aligned}$$

Alors :

$$\mathop{\mathbf{Pr}}_{S\sim(P)^m}\left[\frac{m-|\mathbf{i}_{\max}|}{2B^2}\left(\mathcal{M}_S^{\rho}-\mathcal{M}_P^{\rho}\right)^2\leq\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(P)^m}X_{\pi}\right)\right]\geq 1-\delta.$$

Il ne reste plus qu'à borner le terme : $\mathop{\mathbf{E}}_{S\sim(P)^m}X_{\pi}$. Comme expliqué dans la section 4.1.3, $\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}$ peut contenir un certain biais. Nous considérons alors $\mathcal{M}^{\omega}_{S\setminus S_{\mathbf{i}}}$ la marge empirique définie sur les exemples de l'échantillon d'apprentissage S n'appartenant pas à la séquence de compression $S_{\mathbf{i}}$. $\mathcal{M}^{\omega}_{S\setminus S_{\mathbf{i}}}$ est en fait une moyenne arithmétique de $(m-|\mathbf{i}|)$ variables aléatoires *i.i.d.* et de valeur proche de $\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}$. Concrètement, on a :

$$0\leq m\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-(m-|\mathbf{i}|)\mathcal{M}^{\omega}_{S\setminus S_{\mathbf{i}}}\leq B|\mathbf{i}|,$$

alors :

$$-B|\mathbf{i}|\ \leq\ -|\mathbf{i}|\mathcal{M}^{\omega}_{S\setminus S_{\mathbf{i}}}\ \leq\ m\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-m\mathcal{M}^{\omega}_{S\setminus S_{\mathbf{i}}}\ \leq\ |\mathbf{i}|-|\mathbf{i}|\mathcal{M}^{\omega}_{S\setminus S_{\mathbf{i}}}\ \leq\ B|\mathbf{i}|.$$

Ainsi, on a :

$$\left|\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}^{\omega}_{S\setminus S_{\mathbf{i}}}\right|\leq\frac{B|\mathbf{i}|}{m}. \tag{C.1}$$

Étant donnée une séquence de compression $S_{\mathbf{i}}$, on note $\bar{\mathbf{i}}$ le vecteur des indices n'appartenant pas à $\mathbf{i}$. Alors :

$$\begin{aligned}\mathop{\mathbf{E}}_{S\sim(P)^m}X_{\pi}&=\mathop{\mathbf{E}}_{S\sim(P)^m}\mathop{\mathbf{E}}_{h^{\omega}_{S_{\mathbf{i}}}\sim\pi}\exp\left[m_{\mathbf{i}}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}}}\right)^2\right]\\&=\mathop{\mathbf{E}}_{\mathbf{i}\sim\pi}\mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim(P)^{|\mathbf{i}|}}\mathop{\mathbf{E}}_{\omega\sim\pi_{S_{\mathbf{i}}}}\mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}}\exp\left[m_{\mathbf{i}}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}}}\right)^2\right].\end{aligned}$$

Pour tout $\mathbf{i} \in \mathbf{I}_m$, $S_{\mathbf{i}} \in (X \times Y)^{|\mathbf{i}|}$, $\omega \in \Omega'_{S_{\mathbf{i}}} \times \{+,-\}$, on a :

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi &= \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left[m_{\mathbf{i}}\left(\mathcal{M}_S^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}\right)^2\right] \\
&= \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left[m_{\mathbf{i}}\left(\mathcal{M}_S^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} + \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}\right)^2\right] \\
&\leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left[m_{\mathbf{i}}\left(\left[\mathcal{M}_S^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}}\right]^2 + 2\left|\mathcal{M}_S^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}}\right|\left|\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}\right|\right.\right. \\
&\qquad\qquad \left.\left. + \left[\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}\right]^2\right)\right].
\end{aligned}$$

D'après l'équation (C.1) et puisque $\exp(\cdot)$ est une fonction croissante, on obtient :

$$\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi \leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left[m_{\mathbf{i}}\left(\left[\frac{B|\mathbf{i}|}{m}\right]^2 + 2\frac{B|\mathbf{i}|}{m} + \left[\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}\right]^2\right)\right].$$

Par définition de $\mathbf{i}_{\max}$, pour tout vecteur $\mathbf{i}$ on a : $|\mathbf{i}| \leq |\mathbf{i}_{\max}| \leq m$. Ainsi :

$$\frac{m-|\mathbf{i}|}{2B}\left(\left[\frac{|\mathbf{i}|}{m}\right]^2 + 2\frac{|\mathbf{i}|}{m}\right) \leq |\mathbf{i}_{\max}|\left(\frac{m-|\mathbf{i}|}{2B}\left[\frac{|\mathbf{i}|}{m^2} + \frac{2}{m}\right]\right) \leq \frac{|\mathbf{i}_{\max}|}{B}.$$

Alors :

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi &\leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left[\frac{|\mathbf{i}_{\max}|}{B} + m_{\mathbf{i}}\left(\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}\right)^2\right] \\
&\leq \exp\left[\frac{|\mathbf{i}_{\max}|}{B}\right] + \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left[m_{\mathbf{i}}\left(\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} - \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}\right)^2\right] \\
&\leq \exp\left[\frac{|\mathbf{i}_{\max}|}{B}\right] + \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left[2(m-|\mathbf{i}|)\left(\left[\frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}}}{2B}\right] - \left[\frac{1}{2} - \frac{\mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}}{2B}\right]\right)^2\right].
\end{aligned}$$

Par définition, pour tout $(a,b) \in [0,1]^2$ tels que $a = 0$ impliques $b = 0$ et $a = 1$ implique $b = 1$, on a :

$$\begin{aligned}
2(a-b)^2 &\leq \mathrm{kl}(a\|b) \\
&= a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}.
\end{aligned}$$

Puisque les fonctions de $\mathcal{H}^S$ sont bornées par B et puisque $S_{\bar{\mathbf{i}}}$ est tiré *i.i.d.* selon $(P)^m$, on a :

$$\mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}} = -B \Longrightarrow \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} = -B, \quad \text{et} \quad \mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}} = B \Longrightarrow \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}} = B.$$

Alors :

$$\frac{1}{2} - \frac{\mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}}{2B} = 0 \Longrightarrow \frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}}}{2B} = 0, \quad \text{et} \quad \frac{1}{2} - \frac{\mathcal{M}_P^{h^\omega_{S_{\mathbf{i}}}}}{2B} = 1 \Longrightarrow \frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_{\mathbf{i}}}}}{2B} = 1.$$

De plus, on a :

$$0 \leq \frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega}}{2B} \leq 1, \quad \text{et} \quad 0 \leq \frac{1}{2} - \frac{\mathcal{M}_P^{h_{S_\mathbf{i}}^\omega}}{2B} \leq 1,$$

D'où :

$$\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi \leq \exp\left(\frac{|\mathbf{i}_{\max}|}{B}\right) + \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} \exp\left((m-|\mathbf{i}|)\,\mathrm{kl}\left(\frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega}}{2B} \middle\| \frac{1}{2} - \frac{\mathcal{M}_P^{h_{S_\mathbf{i}}^\omega}}{2B}\right)\right).$$

On applique maintenant le lemme de Maurer (lemme A.1, annexe A) :

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi &\leq \exp\left(\frac{|\mathbf{i}_{\max}|}{B}\right) + \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}|}} 2\sqrt{(m-|\mathbf{i}|)} \\
&\leq \exp\left(\frac{|\mathbf{i}_{\max}|}{B}\right) + 2\sqrt{(m-|\mathbf{i}|)} \\
&\leq \exp\left(\frac{|\mathbf{i}_{\max}|}{B}\right) + 2\sqrt{m}.
\end{aligned}$$

Finalement :

$$\mathop{\mathbf{Pr}}_{S\sim(P)^m}\left(\begin{array}{l} \text{Pour toute distribution } \pi\text{-alignée } \rho \text{ sur } \mathcal{H}^S, \\ \\ |\mathcal{M}_P^\rho - \mathcal{M}_S^\rho| \leq \dfrac{2B\sqrt{\dfrac{|\mathbf{i}_{\max}|}{B\delta} + \ln\left(\dfrac{2\sqrt{m}}{\delta}\right)}}{\sqrt{2(m-|\mathbf{i}_{\max}|)}} \end{array}\right) \geq 1-\delta$$

□

Preuve de l'équation (4.3)**.**

En utilisant des arguments similaires à la preuve précédente. On pose :

$$\begin{aligned}
\mathcal{M}_P^{h_{S_\mathbf{i}}^{(\sigma,+)}, h_{S_{\mathbf{i}'}}^{(\sigma,+)}} &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} h_{S_\mathbf{i}}^{(\sigma,+)}(\mathbf{x}) h_{S_{\mathbf{i}'}}^{(\sigma,+)}(\mathbf{x}), \\
\mathcal{M}_S^{h_{S_\mathbf{i}}^{(\sigma,+)}, h_{S_{\mathbf{i}'}}^{(\sigma,+)}} &= \frac{1}{m}\sum_{i=1}^m h_{S_\mathbf{i}}^{(\sigma,+)}(\mathbf{x}_i) h_{S_{\mathbf{i}'}}^{(\sigma,+)}(\mathbf{x}_i).
\end{aligned}$$

On a :

$$\begin{aligned}
\left(\mathcal{M}_S^{h_{S_\mathbf{i}}^{(\sigma,+)}, h_{S_{\mathbf{i}'}}^{(\sigma,+)}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^{(\sigma,+)}, h_{S_{\mathbf{i}'}}^{(\sigma,+)}}\right)^2 &= \left(\mathcal{M}_S^{h_{S_\mathbf{i}}^{(\sigma,-)}, h_{S_{\mathbf{i}'}}^{(\sigma,+)}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^{(\sigma,-)}, h_{S_{\mathbf{i}'}}^{(\sigma,+)}}\right)^2 \\
&= \left(\mathcal{M}_S^{h_{S_\mathbf{i}}^{(\sigma,+)}, h_{S_{\mathbf{i}'}}^{(\sigma,-)}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^{(\sigma,+)}, h_{S_{\mathbf{i}'}}^{(\sigma,-)}}\right)^2 \\
&= \left(\mathcal{M}_S^{h_{S_{\mathbf{i}'}}^{(\sigma,-)}, h_{S_\mathbf{i}}^{(\sigma,-)}} - \mathcal{M}_P^{h_{S_{\mathbf{i}'}}^{(\sigma,-)}, h_{S_\mathbf{i}}^{(\sigma,-)}}\right)^2.
\end{aligned}$$

Similairement à [McAllester, 2003], nous considérons la transformée de Laplace suivante :

$$X_\pi = \mathop{\mathbf{E}}_{\left(h_{S_\mathbf{i}}^\omega, h_{S_\mathbf{i}}^{\omega}\right)\sim\pi^2} \exp\left[\frac{m-|\mathbf{i}\cup\mathbf{i}'|}{2B^4}\left(\mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}}\right)^2\right].$$

Afin de simplifier la lecture de la preuve nous posons :

$$m_{\mathbf{i}\cup\mathbf{i}'} = \frac{m - |\mathbf{i}\cup\mathbf{i}'|}{2B^4}.$$

La fonction $f(a,b) = \frac{1}{2B^4}(a-b)^2$ est convexe (car sa matrice hessienne est semi-définie positive). Pour toute distribution $\boldsymbol{\pi}$-alignée ρ sur $\mathcal{H}^S$, on a :

$$\begin{aligned}
4X_\pi &= \mathop{\mathbf{E}}_{\left(h^\omega_{S_\mathbf{i}}, h^\omega_{S'_\mathbf{i}}\right)\sim\pi^2} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2\right] \\
&= \int_{\left(h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \pi\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right)\pi\left(h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,+)}_{S_\mathbf{i}} h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) \\
&+ \int_{\left(h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \pi\left(h^{(\sigma,-)}_{S_\mathbf{i}}\right)\pi\left(h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,-)}_{S_\mathbf{i}} h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right) \\
&+ \int_{\left(h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \pi\left(h^{(\sigma,-)}_{S_\mathbf{i}}\right)\pi\left(h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,-)}_{S_\mathbf{i}} h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) \\
&+ \int_{\left(h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \pi\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right)\pi\left(h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,+)}_{S_\mathbf{i}} h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right) \\
&= \int_{\left(h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \left[\pi\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right) + \pi\left(-h^{(\sigma,+)}_{S_\mathbf{i}}\right)\right]\left[\pi\left(h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) + \pi\left(-h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right)\right] \\
&\qquad\qquad \times \exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,+)}_{S_\mathbf{i}} h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) \\
&= \int_{\left(h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \left[\rho\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right) + \rho\left(-h^{(\sigma,+)}_{S_\mathbf{i}}\right)\right]\left[\rho\left(h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) + \rho\left(-h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right)\right] \\
&\qquad\qquad \times \exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,+)}_{S_\mathbf{i}} h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) \\
&= \int_{\left(h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \rho\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right)\rho\left(h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,+)}_{S_\mathbf{i}} h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) \\
&+ \int_{\left(h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \rho\left(h^{(\sigma,-)}_{S_\mathbf{i}}\right)\rho\left(h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,-)}_{S_\mathbf{i}} h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right) \\
&+ \int_{\left(h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \rho\left(h^{(\sigma,-)}_{S_\mathbf{i}}\right)\rho\left(h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,-)}_{S_\mathbf{i}}, h^{(\sigma,+)}_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,-)}_{S_\mathbf{i}} h^{(\sigma,+)}_{S_{\mathbf{i}'}}\right) \\
&+ \int_{\left(h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S'_\mathbf{i}}\right)\in(\mathcal{H}^S)^2} \rho\left(h^{(\sigma,+)}_{S_\mathbf{i}}\right)\rho\left(h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right)\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{(\sigma,+)}_{S_\mathbf{i}}, h^{(\sigma,-)}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^{(\sigma,+)}_{S_\mathbf{i}}, h_{S_{\mathbf{i}'}}}\right)^2\right] d\left(h^{(\sigma,+)}_{S_\mathbf{i}} h^{(\sigma,-)}_{S_{\mathbf{i}'}}\right) \\
&= 4 \mathop{\mathbf{E}}_{\left(h^\omega_{S_\mathbf{i}}, h^\omega_{S'_\mathbf{i}}\right)\sim\rho^2} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}} - \mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2\right] \\
&= 4X_\rho.
\end{aligned}$$

En appliquant l'inégalité de Markov (théorème A.4, annexe A), pour tout $\delta \in (0,1]$ on obtient :

$$\mathop{\mathbf{Pr}}_{S\sim(P)^m}\left(X_\pi \leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi\right) \geq 1-\delta.$$

On applique le logarithme aux deux côtés de l'inégalité précédente. Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P)^m$, pour toute distribution

π-alignée ρ sur $\mathcal{H}^S$, on a :

$$\ln\left(\mathop{\mathbf{E}}_{\left(h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}\right)\sim\rho^2}\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2\right]\right)\leq\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(P)^m}X_\pi\right).$$

L'inégalité de Jensen (théorème A.5, annexe A) à la fonction $\ln(\cdot)$:

$$\ln\left(\mathop{\mathbf{E}}_{\left(h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}\right)\sim\rho^2}\exp\left[m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2\right]\right)\geq\underbrace{\mathop{\mathbf{E}}_{\left(h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}\right)\sim\rho^2}m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2}_{(\mathbf{g}')}.$$

Rappelons que $|\mathbf{i}_{\max}|<\frac{m}{2}$ est la taille maximale de la séquence de compression. Alors en appliquant l'inégalité de Jensen sur $(\mathbf{g}')$ avec la fonction convexe $(m-|\mathbf{i}_{\max}|)f(a,b)=\frac{m-|\mathbf{i}_{\max}|}{2B^4}(a-b)^2\leq m_{\mathbf{i}\cup\mathbf{i}'}(a-b)^2$, on obtient :

$$\begin{aligned}
\mathop{\mathbf{E}}_{\left(h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}\right)\sim\rho^2}m_{\mathbf{i}\cup\mathbf{i}'}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2&=\frac{m}{2B^4}\left(\mathop{\mathbf{E}}_{\left(h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}\right)\sim\rho^2}(-|\mathbf{i}\cup\mathbf{i}'|)\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2\right)\\
&\geq\frac{m-2|\mathbf{i}_{\max}|}{2B^4}\left(\mathop{\mathbf{E}}_{\left(h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}\right)\sim\rho^2}\left(\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-\mathcal{M}_P^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}\right)^2\right)\\
&\geq\frac{m-2|\mathbf{i}_{\max}|}{2B^4}\left(\mathcal{M}_S^{\rho^2}-\mathcal{M}_P^{\rho^2}\right)^2.
\end{aligned}$$

Alors :

$$\mathop{\mathbf{Pr}}_{S\sim(P)^m}\left[\frac{m-2|\mathbf{i}_{\max}|}{2B^4}\left(\mathcal{M}_S^{\rho^2}-\mathcal{M}_P^{\rho^2}\right)^2\leq\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(P)^m}X_\pi\right)\right]\geq1-\delta.$$

Il ne reste plus qu'à borner le terme : $\mathop{\mathbf{E}}_{S\sim(P)^m}X_\pi$. Nous considérons $\mathcal{M}^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}_{S\setminus(S_{\mathbf{i}}\cup S_{\mathbf{i}'})}$ le second moment de la marge empirique défini sur les exemples de l'échantillon d'apprentissage S n'appartenant pas à la séquence de compression $S_{\mathbf{i}}$. $\mathcal{M}^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}_{S\setminus(S_{\mathbf{i}}\cup S_{\mathbf{i}'})}$ est en fait une moyenne arithmétique de $(m-|\mathbf{i}\cup\mathbf{i}'|)$ variables aléatoires *i.i.d.* et de valeur proche de $\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}$. Concrètement, on a :

$$0\ \leq\ m\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-(m-|\mathbf{i}\cup\mathbf{i}'|)\mathcal{M}^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}_{S\setminus(S_{\mathbf{i}}\cup S_{\mathbf{i}'})}\ \leq\ B^2|\mathbf{i}\cup\mathbf{i}'|,$$

alors :

$$\begin{aligned}
-B^2|\mathbf{i}\cup\mathbf{i}'|&\leq-|\mathbf{i}\cup\mathbf{i}'|\mathcal{M}^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}_{S\setminus(S_{\mathbf{i}}\cup S_{\mathbf{i}'})}\\
&\leq m\mathcal{M}_S^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}-m\mathcal{M}^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}_{S\setminus(S_{\mathbf{i}}\cup S_{\mathbf{i}'})}\\
&\leq|\mathbf{i}\cup\mathbf{i}'|-|\mathbf{i}\cup\mathbf{i}'|\mathcal{M}^{h^{\omega}_{S_{\mathbf{i}}},h^{\omega'}_{S_{\mathbf{i}'}}}_{S\setminus(S_{\mathbf{i}}\cup S_{\mathbf{i}'})}\\
&\leq B^2|\mathbf{i}\cup\mathbf{i}'|.
\end{aligned}$$

Donc :

$$\left| \mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_{S\setminus(S_\mathbf{i}\cup S_{\mathbf{i}'})}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right| \leq \frac{B^2|\mathbf{i}\cup\mathbf{i}'|}{m}. \tag{C.2}$$

Étant données deux séquences de compression $S_\mathbf{i}$ et $S_{\mathbf{i}'}$, on pose $\bar{\mathbf{i}}$ le vecteur des indices n'appartenant pas à $\mathbf{i}\cup\mathbf{i}'$. Alors :

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi = \mathop{\mathbf{E}}_{S\sim(P)^m} \mathop{\mathbf{E}}_{\left(h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}\right)\sim\pi^2} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_\mathbf{i}}^\omega} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right)^2 \right] \\
&= \mathop{\mathbf{E}}_{\mathbf{i},\mathbf{i}'\sim\pi^2} \mathop{\mathbf{E}}_{S_\mathbf{i},S_{\mathbf{i}'}\sim(P)^{|\mathbf{i}|}\times(P)^{|\mathbf{i}'|}} \mathop{\mathbf{E}}_{(\omega,\omega')\sim\pi_{S_\mathbf{i}}\times\pi_{S_{\mathbf{i}'}}} \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right)^2 \right].
\end{aligned}$$

On a :

$$\begin{aligned}
&\forall(\mathbf{i},\mathbf{i}')\in(\mathbf{I}_m)^2, \forall(S_\mathbf{i},S_{\mathbf{i}'})\in(X\times Y)^{|\mathbf{i}|}\times(X\times Y)^{|\mathbf{i}'|}, \forall(\omega,\omega')\in(\Omega'_{S_\mathbf{i}}\times\{+,-\})\times(\Omega'_{S_{\mathbf{i}'}}\times\{+,-\}), \\
&\mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^\omega} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right)^2 \right] \\
&= \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^\omega} - \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} + \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right)^2 \right] \\
&\leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\left[\mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right]^2 + 2 \left| \mathcal{M}_S^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right| \left| \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right| \right. \right. \\
&\qquad\qquad \left. \left. + \left[\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right]^2 \right) \right].
\end{aligned}$$

D'après l'équation (C.2) et puisque $\exp(\cdot)$ est une fonction croissante, on a :

$$\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi \leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\left[\frac{B^2|\mathbf{i}\cup\mathbf{i}'|}{m} \right]^2 + 2\frac{B^2|\mathbf{i}\cup\mathbf{i}'|}{m} + \left[\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right]^2 \right) \right].$$

Pour tout $\mathbf{i}$, on suppose que $|\mathbf{i}| \leq |\mathbf{i}_{\max}| \leq \frac{m}{2}$. Donc :

$$m_{\mathbf{i}\cup\mathbf{i}'} \left(\left[\frac{|\mathbf{i}\cup\mathbf{i}'|}{m} \right]^2 + 2\frac{|\mathbf{i}\cup\mathbf{i}'|}{m} \right) \leq 2|\mathbf{i}_{\max}| \left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\frac{|\mathbf{i}\cup\mathbf{i}'|}{m^2} + \frac{2}{m} \right) \right] \leq \frac{2|\mathbf{i}_{\max}|}{B^2}.$$

Alors :

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi &\leq \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[\frac{2|\mathbf{i}_{\max}|}{B^2} + m_{\mathbf{i}\cup\mathbf{i}'} \left(\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right)^2 \right] \\
&\leq \exp\left[\frac{2|\mathbf{i}_{\max}|}{B^2} \right] + \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[m_{\mathbf{i}\cup\mathbf{i}'} \left(\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} - \mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}} \right)^2 \right] \\
&\leq \exp\left[\frac{2|\mathbf{i}_{\max}|}{B^2} \right] \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[2(m-|\mathbf{i}\cup\mathbf{i}'|) \left(\left[\frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}}}{2B} \right] - \left[\frac{1}{2} - \frac{\mathcal{M}_P^{h_{S_\mathbf{i}}^\omega, h_{S_{\mathbf{i}'}}^{\omega'}}}{2B} \right] \right)^2 \right].
\end{aligned}$$

Rappelons que $2(a-b)^2 \leq \mathrm{kl}(a\|b)$ est vrai pour tout $(a,b) \in [0,1]^2$ tels que si $a=0$ alors $b=0$ et si $a=1$ alors $b=1$. Puisque les fonctions de $\mathcal{H}^S$ sont bornées par B et $S_{\bar{\mathbf{i}}}$ est *i.i.d.* selon D, on a :

$$\mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}} = -B^2 \Rightarrow \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}} = -B^2, \quad \text{et} \quad \mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}} = B^2 \Rightarrow \mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}} = B^2.$$

Alors :

$$\frac{1}{2} - \frac{\mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2} = 0 \Rightarrow \frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2} = 0, \quad \text{et} \quad \frac{1}{2} - \frac{\mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2} = 1 \Rightarrow \frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2} = 1.$$

Comme :

$$0 \leq \frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2} \leq 1, \quad \text{et} \quad 0 \leq \frac{1}{2} - \frac{\mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2} \leq 1,$$

on a :

$$\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi \leq \exp\left[\frac{2|\mathbf{i}_{\max}|}{B^2}\right] + \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} \exp\left[(m-|\mathbf{i}\cup\mathbf{i}'|)\,\mathrm{kl}\left(\frac{1}{2} - \frac{\mathcal{M}_{S_{\bar{\mathbf{i}}}}^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2} \middle\| \frac{1}{2} - \frac{\mathcal{M}_P^{h^\omega_{S_\mathbf{i}}, h^{\omega'}_{S_{\mathbf{i}'}}}}{2B^2}\right)\right].$$

En appliquant le lemme de Maurer (lemme A.1, annexe A), on obtient :

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim(P)^m} X_\pi &\leq \exp\left(\frac{2|\mathbf{i}_{\max}|}{B^2}\right) + \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim(P)^{m-|\mathbf{i}\cup\mathbf{i}'|}} 2\sqrt{(m-|\mathbf{i}\cup\mathbf{i}'|)} \\
&\leq \exp\left(\frac{2|\mathbf{i}_{\max}|}{B^2}\right) + 2\sqrt{(m-|\mathbf{i}\cup\mathbf{i}'|)} \\
&\leq \exp\left(\frac{2|\mathbf{i}_{\max}|}{B^2}\right) + 2\sqrt{m}.
\end{aligned}$$

Finalement :

$$\mathop{\mathbf{Pr}}_{S\sim(P)^m}\left(\begin{array}{l} \text{pour toute distribution } \pi\text{-alignée } \rho \text{ sur } \mathcal{H}^S, \\ \left|\mathcal{M}_P^{\rho^2} - \mathcal{M}\rho^2{}_S\right| \leq \dfrac{2B^2\sqrt{\dfrac{2|\mathbf{i}_{\max}|}{B^2\delta} + \ln\left(\dfrac{2\sqrt{m}}{\delta}\right)}}{\sqrt{2(m-2|\mathbf{i}_{\max}|)}} \end{array}\right) \geq 1-\delta.$$

Annexe du chapitre 5

D

D.1 Preuves de la proposition 5.1 et de son corollaire 5.3

D.1.1 Preuve de la proposition 5.1

Soit un exemple étiqueté $(\mathbf{x}, y)$. Pour toute classe $c \in Y$, on définit $\gamma_c(\mathbf{x})$ tel que :

$$\gamma_c(\mathbf{x}) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) = \sum_{h\in\mathcal{H}:h(\mathbf{x})=c} \rho(h).$$

Bien entendu, on a :

$$\sum_{c\in Y} \gamma_c(\mathbf{x}) = 1.$$

Rappelons que les risques conditionnels de Gibbs $R(G_\rho, p, q)$ et du vote de majorité $R(B_\rho, p, q)$ sont définis par :

$$R(G_\rho, p, q) = \mathop{\mathbf{E}}_{\mathbf{x}\sim P_{|y=p}} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}\left(h(\mathbf{x}) = q\right) = \mathop{\mathbf{E}}_{\mathbf{x}\sim D_{|y=p}} \gamma_q(\mathbf{x}), \tag{5.16}$$

$$R(B_\rho, p, q) = \mathop{\mathbf{E}}_{\mathbf{x}\sim P_{|y=p}} \mathbf{I}\left(\operatorname*{argmax}_{c\in Y} \gamma_c(\mathbf{x}) = q\right), \tag{5.17}$$

où $P_{|y=p}$ est la probabilité conditionnelle d'un exemple $\mathbf{x}$ sachant la classe y. L'équation (5.16) correspond à l'élément d'indice (p, q) dans la matrice $\mathbf{C}_P^{G_\rho}$ (si $p \neq q$) et l'équation (5.17) à l'élément (p, q) de $\mathbf{C}_P^{B_\rho}$.
Pour qu'une classe $q \neq y$ soit prédite par le vote de majorité $B_\rho(\cdot)$, il est nécessaire et suffisant que :

$$\forall c \in Y, c \neq q,\ \gamma_q(\mathbf{x}) \geq \gamma_c(\mathbf{x}).$$

Ce qui est équivalent à :

$$\mathbf{I}\left(\operatorname*{argmax}_{c\in Y} \gamma_c(\mathbf{x}) = q\right) = \mathbf{I}\left(\bigwedge_{c\in Y:c\neq q} \gamma_q(\mathbf{x}) \geq \gamma_c(\mathbf{x})\right). \tag{D.1}$$

Notons que l'espérance selon $P_{|y=p}$ du côté gauche de l'égalité correspond à $R(B_\rho, p, q)$ — voir l'équation (5.17). On peut maintenant remarquer :

$$\begin{aligned}
\mathbf{I}\left(\bigwedge_{c\in Y:c\neq q} \gamma_q(\mathbf{x}) \geq \gamma_c(\mathbf{x})\right) = 1 &\Leftrightarrow \forall c \in Y,\ c \neq q,\ \gamma_q(\mathbf{x}) - \gamma_c(\mathbf{x}) \geq 0 \\
&\Rightarrow \sum_{c\in Y:c\neq q} (\gamma_q(\mathbf{x}) - \gamma_c(\mathbf{x})) \geq 0 \\
&\Leftrightarrow \sum_{c\in Y:c\neq q} \gamma_q(\mathbf{x}) - \sum_{c\in Y:c\neq q} \gamma_c(\mathbf{x}) \geq 0 \\
&\Leftrightarrow (Q-1)\gamma_q(\mathbf{x}) - (1-\gamma_q(\mathbf{x})) \geq 0 \\
&\Leftrightarrow \gamma_q(\mathbf{x}) \geq \frac{1}{Q}.
\end{aligned}$$

Les deux dernières lignes sont dûes à $\sum_{c\in Y} \gamma_c(\mathbf{x}) = 1$. Ce résultat signifie que :

$$\mathbf{I}\left(\bigwedge_{c\in Y:c\neq q} \gamma_q(\mathbf{x}) \geq \gamma_c(\mathbf{x})\right) = 1 \Rightarrow \mathbf{I}\left(\gamma_q(\mathbf{x}) \geq \frac{1}{Q}\right) = 1,$$

impliquant :

$$\mathbf{I}\left(\bigwedge_{c\in Y:c\neq q} \gamma_q(\mathbf{x}) \geq \gamma_c(\mathbf{x})\right) \leq \mathbf{I}\left(\gamma_q(\mathbf{x}) \geq \frac{1}{Q}\right),$$

ainsi, grâce à (D.1) :

$$\mathbf{I}\left(\operatorname*{argmax}_{c\in Y} \gamma_c(\mathbf{x}) = q\right) \leq \mathbf{I}\left(\gamma_q(\mathbf{x}) \geq \frac{1}{Q}\right). \tag{D.2}$$

Nous utilisons ensuite : $\forall \gamma \in [0,1], \theta \in [0,1],\ \gamma \geq \theta \mathbf{I}\left(\gamma \geq \frac{1}{Q}\right)$ (illustré sur la figure D.1), pour obtenir :

$$\frac{1}{Q}\,\mathbf{I}\left(\gamma_q(\mathbf{x}) \geq \frac{1}{Q}\right) \leq \gamma_q(\mathbf{x}) \Leftrightarrow \mathbf{I}\left(\gamma_q(\mathbf{x}) \geq \frac{1}{Q}\right) \leq Q\gamma_q(\mathbf{x}).$$

En combinant ce résultat avec l'inégalité (D.2), on a :

$$\mathbf{I}\left(\operatorname*{argmax}_{c\in Y} \gamma_c(\mathbf{x}) = q\right) \leq Q\gamma_q(\mathbf{x}).$$

En prenant l'espérance $D_{|y=p}$ des deux côtés de l'inégalité, on obtient :

$$R(B_\rho, p, q) \leq QR(G_\rho, p, q).$$

D.1.2 Preuve du corollaire 5.3

D'après les définitions de $R(G_\rho, p, q)$ de l'équation (5.16) et $R(B_\rho, p, q)$ de l'équation (5.17), la proposition 5.1, implique directement :

$$\mathbf{C}^{B_\rho} \leq Q\mathbf{C}^{G_\rho}, \tag{D.3}$$

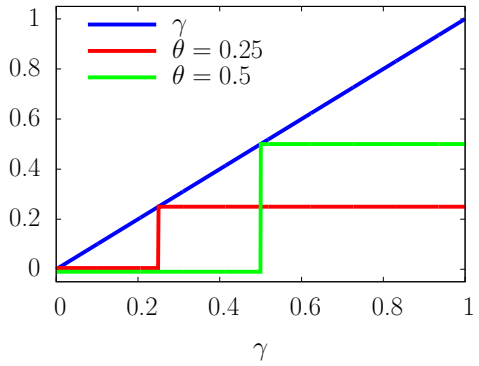


FIGURE D.1 – *Graphe de* $\gamma \mapsto \theta \mathbf{I}(\gamma \geq \theta)$ *, pour* $\theta = 0.25$ *(rouge) et* $\theta = 0.5$ *(vert). On observe que :* $\forall \theta \in [0,1]$, $\gamma \geq \theta \mathbf{I}(\gamma \geq \theta)$.

où nous rappelons que $\leq$ est la relation d'ordre sur les matrices élément par élément. En considèrant les matrices de dilatation de $\mathbf{C}^{B_\rho}$ et $Q\mathbf{C}^{G_\rho}$, l'équation (D.3) s'écrit :

$$\mathcal{D}(\mathbf{C}^{B_\rho}) \leq \mathcal{D}(Q\mathbf{C}^{G_\rho}).$$

Puisque tous les éléments d'une matrice de confusion sont positifs, on a :

$$0 \leq \mathcal{D}(\mathbf{C}^{B_\rho}) \leq \mathcal{D}(Q\mathbf{C}^{G_\rho}).$$

En appliquant l'équation (5.4), on obtient :

$$\lambda_{\max}(\mathcal{D}(\mathbf{C}^{B_\rho})) \leq \lambda_{\max}(\mathcal{D}(Q\mathbf{C}^{G_\rho})). \tag{D.4}$$

Alors, l'équation (5.6) appliquée à l'équation (D.4) donne :

$$\left\|\mathbf{C}^{B_\rho}\right\| \leq \left\|Q\mathbf{C}^{G_\rho}\right\|.$$

Finalement, l'équation (5.3) implique :

$$\left\|\mathbf{C}^{B_\rho}\right\| \leq Q\left\|\mathbf{C}^{G_\rho}\right\|.$$

D.2 Preuve du théorème 5.4

Selon le même principe que la preuve de la C-borne en annexe B.3, on fait appel à l'inégalité de Cantelli-Chebychev (théorème A.2) en remplaçant la variable Z par la variable aléatoire $\mathcal{M}^\rho(\mathbf{x}, y)$ et a par $\mathcal{M}_P^\rho$. Rappelons que d'après notre définition, on a : $\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x}, y) = \mathcal{M}_P^{\rho^2} - (\mathcal{M}_P^\rho)^2$. Ainsi :

$$\underset{(\mathbf{x},y)\sim P}{\mathbf{Pr}}\ (\mathcal{M}^\rho(\mathbf{x}, y) \leq 0) \leq 1 - \frac{\left(\mathcal{M}_P^\rho\right)^2}{\mathcal{M}_P^{\rho^2}} = C_P^\rho.$$

D.3 Preuve du théorème 5.6

D'après la définition 5.5 de $\ell_P^\rho(\omega)$, on a :

$$\begin{aligned}\ell_P^\rho(\omega) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)\leq\omega\right]\\ &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\leq 0\right]. \qquad \text{(D.5)}\end{aligned}$$

On applique l'inégalité de Cantelli-Chebitchev sur la ligne (D.5), en remplaçant Z par la variable aléatoire $\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)$ et a par $\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right]$. On obtient :

$$\ell_P^\rho(\omega)\leq\frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)^2}.$$

Finalement, puisque :

$$\begin{aligned}&\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)\\ =\ &\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)^2-\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)\right]^2,\end{aligned}$$

on a :

$$\ell_P^\rho(\omega)\leq 1-\frac{\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)\right]^2}{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)-\omega\right)^2}.$$

D.4 Preuve du théorème 5.3

Tout d'abord, nous prouvons la partie gauche de l'inégalité (5.25) :

$$\ell_P^\rho(\tfrac{1}{Q}) \leq \mathbf{R}_P(B_\rho).$$

En fait, on a :

$$\begin{aligned}
\mathbf{R}_P(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathcal{M}^\rho(\mathbf{x},y) \leq 0\right] \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\exists c \in Y, c \neq y : \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right] \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \max_{c\in Y, c\neq y}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right] \\
&\geq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \frac{1-\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)}{Q-1}\right] \qquad \text{(D.6)} \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \frac{1}{Q}\right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\mathbf{I}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \frac{1}{Q}\right] \\
&= \ell_P^\rho(\tfrac{1}{Q}).
\end{aligned}$$

La ligne (D.6) provient de : $\dfrac{1-\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y)}{Q-1} \leq \max_{c\in Y, c\neq y}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)$.

Ensuite, nous prouvons la partie droite de l'inégalité (5.25) : $\mathbf{R}_P(B_\rho) \leq \ell_P^\rho(\omega)\frac{1}{2}$. L'assertion précédente est vérifiée si :

$$\begin{aligned}
R(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathcal{M}^\rho(\mathbf{x},y) \leq 0\right] \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \max_{c\in Y, c\neq y}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right] \\
&\leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \frac{1}{2}\right] \\
&= \ell_P^\rho(\tfrac{1}{2}).
\end{aligned}$$

Ce qui est équivalent à vérifier que :

$$\forall(\mathbf{x},y)\sim P,\ \max_{c\in Y, c\neq y}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c) \leq \frac{1}{2}.$$

(i) Selon la définition de $B_\rho(\cdot)$, si pour un exemple $(\mathbf{x},y)$ tiré selon P on a : $\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq 1/2$, alors $\mathcal{M}^\rho(\mathbf{x},y) \leq 0$.

(ii) De plus, d'après la définition de la ω-perte, pour $\omega = \frac{1}{2}$, on a :

$$\begin{aligned}
\mathbf{R}_P(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \max_{c\in Y, c\neq y}\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right] \\
&\leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) \leq \frac{1}{2}\right].
\end{aligned}$$

Le résultat vient de (i) et (ii).

Annexe du chapitre 6

E

E.1 Preuve du lemme 6.1

Rappelons que $F(\cdot)$ fait référence au problème (6.2) de DASF. Pour toute solution $\boldsymbol{\alpha}$, on a :

$$\begin{aligned}\sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \left\| \left(\boldsymbol{\phi}^R(\mathbf{x}_s)^\top - \boldsymbol{\phi}^R(\mathbf{x}_t)^\top\right) \operatorname{diag}(\boldsymbol{\alpha}) \right\|_1 &= \sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \sum_{j=1}^{r} \left| \alpha_j \left(K(\mathbf{x}_s,\mathbf{x}'_j) - K(\mathbf{x}_t,\mathbf{x}'_j) \right) \right| \\ &= \sum_{j=1}^{r} \left[|\alpha_j| \left(\sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \left| K(\mathbf{x}_s,\mathbf{x}'_j) - K(\mathbf{x}_t,\mathbf{x}'_j) \right| \right) \right] \\ &\geq \sum_{j=1}^{r} \left[|\alpha_j| \max_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \left| K(\mathbf{x}_s,\mathbf{x}'_j) - K(\mathbf{x}_t,\mathbf{x}'_j) \right| \right].\end{aligned}$$

D'après l'hypothèse (6.3) et d'après la définition de B_R on a :

$$B_R = \min_{\mathbf{x}'_j\in R} \left\{ \max_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \left| K(\mathbf{x}_s,\mathbf{x}'_j) - K(\mathbf{x}_t,\mathbf{x}'_j) \right| \right\} > 0.$$

Donc :

$$\sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \left\| \left(\boldsymbol{\phi}^R(\mathbf{x}_s)^\top - \boldsymbol{\phi}^R(\mathbf{x}_t)^\top\right) \operatorname{diag}(\boldsymbol{\alpha}) \right\|_1 \geq \|\boldsymbol{\alpha}\|_1 B_R.$$

Alors :

$$\|\boldsymbol{\alpha}^*\|_1(\lambda + \beta B_R) + \frac{1}{m^s} \sum_{i=1}^{m^s} \left[1 - y_i^s \sum_{j=1}^{r} \alpha_j^* K(\mathbf{x}_i^s,\mathbf{x}'_j) \right]_+ \leq F(\boldsymbol{\alpha}^*),$$

où $\boldsymbol{\alpha}^*$ est la solution optimale de $F(\cdot)$. Ainsi, on a :

$$F(\boldsymbol{\alpha}^*) \leq F(\mathbf{0}) = 1,$$

où $\mathbf{0}$ est le vecteur nul. Il en découle :

$$\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\beta B_R + \lambda}.$$

E.2 Preuve du théorème 6.1

Soit (X, ϱ) un espace métrique compact. Soit $\eta > 0$, puisque X est compact, d'après la définition du nombre de couvertures, on partitionne X en M_η sous-ensemble (M_η

fini), tels que pour $\mathbf{x}_1$ et $\mathbf{x}_2$ appartenant au même sous-ensemble, on ait : $\varrho(\mathbf{x}_1, \mathbf{x}_2) \leq \eta$. On divise Y en 2 sous-ensembles $\{\{-1\}, \{+1\}\}$. En suivant le principe de preuve de [Xu et Mannor, 2010], on partitionne $X \times Y$ en $2M_\eta$ sous-ensembles tels que les points appartenant à un même sous-ensemble soient de même classe. Étant donné une fonction de similarité $K(\cdot, \cdot)$ continue sur son premier argument, un échantillon d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s}$ *i.i.d.* selon P_S, un ensemble de *landmarks* $R = \{\mathbf{x}'_j\}_{j=1}^{r}$, les hyperparamètres $\lambda > 0$, $\beta > 0$ et un ensemble de couples $\mathcal{C}_{ST}$, on note $\boldsymbol{\alpha}^*$ la solution optimale du problème (6.2). Pour tout $\mathbf{s}_1 = (\mathbf{x}_1, y_1) \in S$, et pour tout $\mathbf{s}_2 = (\mathbf{x}_2, y_2)$ tels que $\mathbf{s}_1$ et $\mathbf{s}_2$ appartiennent au même sous-ensemble, on a : $y_1 = y_2 = y$ et $\varrho(\mathbf{x}_1, \mathbf{x}_2) \leq \eta$. Alors :

$$\left|\ell_{hinge}(h, (\mathbf{x}_1, y)) - \ell_{hinge}(h, (\mathbf{x}_2, y))\right| = \left| \left[1 - y \sum_{j=1}^{r} \alpha_j^* K(\mathbf{x}_1, \mathbf{x}'_j)\right]_+ - \left[1 - y \sum_{j=1}^{r} \alpha_j^* K(\mathbf{x}_2, \mathbf{x}'_j)\right]_+ \right|.$$

D'après la 1-lipschitznesse de la perte hinge et par l'application de l'inégalité de Hölder (théorème A.3, annexe A), on obtient :

$$\begin{aligned}
\left|\ell_{hinge}(h, (\mathbf{x}_1, y)) - \ell_{hinge}(h, (\mathbf{x}_2, y))\right| &\leq \|\boldsymbol{\alpha}^*\|_1 \left\| \phi^R(\mathbf{x}_1)^\top - \phi^R(\mathbf{x}_2)^\top \right\|_\infty \\
&\leq \|\boldsymbol{\alpha}^*\|_1 \max_{\substack{(\mathbf{x}_a, \mathbf{x}_b) \sim (D_S)^2 \\ \varrho(\mathbf{x}_a, \mathbf{x}_b) \leq \eta}} \left\{ \left\| \phi^R(\mathbf{x}_a)^\top - \phi^R(\mathbf{x}_b)^\top \right\|_\infty \right\} \\
&\leq \frac{N_\eta}{\beta B_R + \lambda},
\end{aligned}$$

avec :

$$N_\eta = \max_{\substack{(\mathbf{x}_a, \mathbf{x}_b) \sim (D_S)^2 \\ \varrho(\mathbf{x}_a, \mathbf{x}_b) \leq \eta}} \left\{ \left\| \phi^R(\mathbf{x}_a)^\top - \phi^R(\mathbf{x}_b)^\top \right\|_\infty \right\},$$

fini par continuité de $K(\cdot, \cdot)$ sur son premier argument et par la définition du nombre de couverture. Alors l'algorithme associé au problème de minimisation (6.2) est $\left(2M_\eta, \frac{N_\eta}{\beta B_R + \lambda}\right)$ robuste sur P_S.

Annexe du chapitre 7

F

F.1 Preuve du théorème 7.1

Considérons la variable aléatoire non négative :

$$\mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m_u(\mathbf{R}_{D_S}(h,h')-\mathbf{R}_{S_u}(h,h'))^2}.$$

En appliquant l'inégalité de Markov (théorème A.4, annexe A). Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u \sim (D_S)^{m_u}$, on a :

$$\begin{aligned}&\mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\\ &\leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S_u\sim(D_S)^{m_u}} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right).\end{aligned}$$

En prenant le logarithme de chaque côté de l'inégalité précédente et en passant de l'espérance selon π^2 à l'espérance selon ρ^2, on obtient, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u \sim (D_S)^{m_u}$, et pour toute distribution posterior ρ :

$$\begin{aligned}&\ln\left[\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \frac{\pi(h)\pi(h')}{\rho(h)\rho(h')} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\right]\\ \leq &\ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(D_S)^{m_u}} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\right].\end{aligned}$$

Puisque $\ln(\cdot)$ est une fonction concave, on peut appliquer l'inégalité de Jensen (théorème A.5, annexe A). Alors, on a :

$$\begin{aligned}&\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \ln\left[\frac{\pi(h)\pi(h')}{\rho(h)\rho(h')} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\right]\\ \leq &\ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S_u} \sim (D_S)^{m_u} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left((2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\right].\end{aligned}$$

Notons que :

$$\begin{aligned}\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \ln\frac{\rho(h)\rho(h')}{\pi(h)\pi(h')} &= \mathop{\mathbf{E}}_{h\sim\rho} \ln\frac{\rho(h)}{\pi(h)} + \mathop{\mathbf{E}}_{h'\sim\rho} \ln\frac{\rho(h')}{\pi(h')}\\ &= 2\mathop{\mathbf{E}}_{h\sim\rho} \ln\frac{\rho(h)}{\pi(h)}\\ &= 2\,\mathrm{KL}(\rho\|\pi).\end{aligned} \tag{F.1}$$

Et donc :

$$\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \ln \frac{\pi(h)\pi(h')}{\rho(h)\rho(h')} = -2\,\mathrm{KL}(\rho\|\pi).$$

Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u \sim (D_S)^{m_u}$, et pour toute distribution posterior ρ, on a :

$$\begin{aligned}
&-2\,\mathrm{KL}(\rho\|\pi) + \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} m_u\, 2(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2 \\
&\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S_u\sim(D_S)^{m_u}} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\right].
\end{aligned}$$

Puisque $\mathcal{D}(a,b) = 2(a-b)^2$ est une fonction convexe, l'inégalité de Jensen implique :

$$\begin{aligned}
&\left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h')\right)\right)^2 \\
&\leq \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} (\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2.
\end{aligned}$$

Ainsi, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u \sim (D_S)^{m_u}$ et pour toute distribution posterior ρ, on a :

$$\begin{aligned}
&2m_u\left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{D_S}(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho^2} \mathbf{R}_{S_u}(h,h')\right)^2 \leq 2\,\mathrm{KL}(\rho\|\pi) \\
&+\ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(D_S)^{m_u}} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\right].
\end{aligned}$$

Bornons maintenant :

$$\ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(D_S)^{m_u}} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left(2m(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right)\right].$$

Pour ce faire, on a :

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S_u\sim(D_S)^{m_u}} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right) \\
&= \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \mathop{\mathbf{E}}_{S_u\sim(D_S)^{m_u}} \exp\left(2m_u(\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{S_u}(h,h'))^2\right) && \text{(F.2)} \\
&\leq \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \mathop{\mathbf{E}}_{S_u\sim(D_S)^{m_u}} \exp\left(\mathrm{kl}(\mathbf{R}_{S_u}(h,h')\|\mathbf{R}_{D_S}(h,h'))\right) && \text{(F.3)} \\
&\leq 2\sqrt{m_u}. && \text{(F.4)}
\end{aligned}$$

La ligne (F.2) provient de l'indépendance entre D_S et π^2. L'inégalité de Pinsker

$$2(q-p)^2 \leq \mathrm{kl}(q\|p) \quad \text{pour tout } (p,q) \in [0,1]^2,$$

implique la ligne (F.3). La dernière ligne (F.4) provient du lemme de Maurer (lemme A.1, annexe A). Donc pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u \sim (D_S)^{m_u}$ et pour toute distribution posterior ρ, on a :

$$\begin{aligned}
&2m\left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{D_S}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{S_u}(h,h')\right)^2 \\
&\leq\ 2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m_u}}{\delta}
\end{aligned}$$

$$\Leftrightarrow \quad \left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{D_S}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{S_u}(h,h') \right)^2 \leq \frac{1}{2m_u} \left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m_u}}{\delta} \right]$$

$$\Leftrightarrow \quad \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{D_S}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{S_u}(h,h') \right| \leq \sqrt{\frac{1}{2m_u} \left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m_u}}{\delta} \right]}. \tag{F.5}$$

De la même manière, on borne :

$$\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{D_T}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{T_u}(h,h') \right|.$$

Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $T_u \sim (D_T)^{m_u}$ et pour toute distribution posterior ρ, on obtient :

$$\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{D_T}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{R}_{T_u}(h,h') \right| \leq \sqrt{\frac{1}{2m_u} \left[2\,\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m_u}}{\delta} \right]}. \tag{F.6}$$

Finalement en remplaçant δ par $\frac{\delta}{2}$ dans les inégalités (F.5) et (F.6). La borne de l'union implique le résultat du théorème, puisque :

$$\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{D_T}(h,h') - \mathbf{R}_{D_S}(h,h') \right] \right| = \mathrm{dis}_\rho(D_S, D_T),$$

$$\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{T_u}(h,h') - \mathbf{R}_{S_u}(h,h') \right] \right| = \mathrm{dis}_\rho(S_u, T_u),$$

et puisque si $|a_1 - b_1| \leq c_1$ et $|a_2 - b_2| \leq c_2$, alors $|(a_1 - a_2) - (b_1 - b_2)| \leq c_1 + c_2$.

F.2 Preuve du théorème 7.2

Nous proposons de borner :

$$d^{(1)} = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{D_T}(h,h') \right],$$

par son estimation empirique :

$$d^{(1)}_{S_u \times T_u} = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{S_u}(h,h') - \mathbf{R}_{T_u}(h,h') \right],$$

et le terme lié à la KL-divergence entre le posterior et le prior.

Pour ce faire, nous considérons le classifieur "abstrait" $\widehat{h} = (h,h') \in \mathcal{H}^2$ choisi aléatoirement selon la distribution $\widehat{\rho}$, où $\widehat{\rho}(\widehat{h}) = \rho(h)\rho(h')$. Signalons qu'avec $\widehat{\pi}(\widehat{h}) = \pi(h)\pi(h')$, on obtient avec l'équation (F.1) : $\mathrm{KL}(\widehat{\rho}\|\widehat{\pi}) = 2\,\mathrm{KL}(\rho\|\pi)$. Nous définissons la fonction de perte "abstraite" pour $\widehat{h}$ et pour une paire d'exemples $(\mathbf{x}^s, \mathbf{x}^t) \sim (D_S \times D_T)$ par :

$$\ell_{d^{(1)}}(\widehat{h}(\mathbf{x}^s), \widehat{h}(\mathbf{x}^t)) = \frac{1 + \ell_{0\text{–}1}(h(\mathbf{x}^s), h'(\mathbf{x}^s)) - \ell_{0\text{–}1}(h(\mathbf{x}^t), h'(\mathbf{x}^t))}{2}.$$

$\ell_{d^{(1)}}(\cdot,\cdot)$ renvoie des valeurs dans $[0,1]$. Ainsi, le risque "abstrait" de $\widehat{h}$ sur la distribution jointe défini par :

$$R^{(1)}_{D_S\times D_T}(\widehat{h}) = \mathop{\mathbf{E}}_{\mathbf{x}^s\sim D_S}\ \mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T} \ell_{d^{(1)}}(\widehat{h}(\mathbf{x}^s),\widehat{h}(\mathbf{x}^t))\,,$$

et le risque du classifieur de Gibbs associé est :

$$R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}}) = \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{D_S\times D_T}(\widehat{h})\,.$$

Les valeurs empiriques de ces deux quantités sont :

$$R^{(1)}_{S_u\times T_u}(\widehat{h}) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x}^s,\mathbf{x}^t)\sim S_u\times T_u} \ell_{d^{(1)}}(\widehat{h}(\mathbf{x}^s),\widehat{h}(\mathbf{x}^t))$$

et,

$$R^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}}) \;=\; \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{S_u\times T_u}(\widehat{h})\,.$$

Considérons la variable aléatoire non négative :

$$\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\,\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\middle\|R^{(1)}_{D_S\times D_T}(\widehat{h})\right)\right).$$

On applique l'inégalité de Markov (théorème A.4, annexe A). Pour tout $\delta\in(0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix $S_u\times T_u\sim(D_S\times D_T)^{m_u}$, on a :

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\,\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\middle\|R^{(1)}_{D_S\times D_T}(\widehat{h})\right)\right)\\
&\le\ \frac{1}{\delta}\mathop{\mathbf{E}}_{S_u\times T_u\sim(D_S\times D_T)^{m_u}}\ \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\,\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\middle\|R^{(1)}_{D_S\times D_T}(\widehat{h})\right)\right).
\end{aligned}$$

En prenant le logarithme de chaque côté de l'inégalité précédente, et en passant de l'espérance selon $\widehat{\pi}$ à l'espérance selon $\widehat{\rho}$, on obtient :

$$\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}}\frac{\widehat{\pi}(\widehat{h})}{\widehat{\rho}(\widehat{h})}\ \exp\left(m_u\,\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\middle\|R^{(1)}_{D_S\times D_T}(\widehat{h})\right)\right)\right] \qquad\text{(F.7)}\\
&\le\ \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S_u\times T_u\sim(D_S\times D_T)^{m_u}}\ \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\,\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\middle\|R^{(1)}_{D_S\times D_T}(\widehat{h})\right)\right)\right]\\
&\le \ln\frac{2\sqrt{m_u}}{\delta}\,.
\end{aligned}$$

La dernière inégalité provient du lemme de Maurer (lemme A.1, annexe A).

Nous dérivons une minoration en faisant appel deux fois à l'inégalité de Jensen (théorème A.5, annexe A) avec la fonction logarithme, puis avec la fonction convexe $\mathrm{kl}(\cdot\|\cdot)$:

$$\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}}\frac{\widehat{\pi}(\widehat{h})}{\widehat{\rho}(\widehat{h})}\ \exp\left(m_u\,\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\middle\|R^{(1)}_{D_S\times D_T}(\widehat{h})\right)\right)\right]\\
&= \ln\left[\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} \exp\left(m_u\,\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\middle\|R^{(1)}_{D_S\times D_T}(\widehat{h})\right)\right)\right] - 2\,\mathrm{KL}(\rho\|\pi)\\
&\ge\ \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} m_u\ \mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\ \middle\|\ R^{(1)}_{D_S\times D_T}(\widehat{h})\right) - 2\,\mathrm{KL}(\rho\|\pi)\\
&\ge\ m_u\ \mathrm{kl}\left(\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{S_u\times T_u}(\widehat{h})\ \middle\|\ \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{D_S\times D_T}(\widehat{h})\right) - 2\,\mathrm{KL}(\rho\|\pi)\\
&\ge\ m_u\ \mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}})\ \middle\|\ R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}})\right) - 2\,\mathrm{KL}(\rho\|\pi)\,.
\end{aligned}$$

Ce qui implique que :

$$\mathrm{kl}\left(R^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}})\middle\|R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}})\right)\leq\frac{1}{m_u}\left[2\,\mathrm{KL}(\rho\,\|\,\pi)+\ln\frac{2\sqrt{m_u}}{\delta}\right].$$

Puisque que pour $d^{(1)}$, on a : $d^{(1)}=2R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}})-1$ et $d^{(1)}_{S_u\times T_u}=2R^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}})-1$, la ligne précédente implique directement une borne pour $d^{(1)}$. Ainsi, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_u\times T_u\sim(D_S\times D_T)^{m_u}$, on a :

$$\mathrm{kl}\left(\frac{d^{(1)}_{S_u\times T_u}+1}{2}\middle\|\frac{d^{(1)}+1}{2}\right)\leq\frac{1}{m_u}\left[2\,\mathrm{KL}(\rho\,\|\,\pi)+\ln\frac{2\sqrt{m_u}}{\delta}\right]. \tag{F.8}$$

De plus, on a :

$$\mathrm{kl}\left(\frac{|d^{(1)}_{S_u\times T_u}|+1}{2}\middle\|\frac{|d^{(1)}|+1}{2}\right)\leq\frac{1}{m}\left[2\,\mathrm{KL}(\rho\,\|\,\pi)+\ln\frac{2\sqrt{m_u}}{\delta}\right], \tag{F.9}$$

puisque :

$$|d^{(1)}|=\mathrm{dis}_\rho(D_S,D_T)\ \text{ et }\ |d^{(1)}_{S_u\times T_u}|=\mathrm{dis}_\rho(S_u,T_u)\,.$$

Pour finaliser la preuve, prouvons l'équation (F.9). Quatre cas sont à considérer.

1 : $d^{(1)}_{S_u\times T_u}\geq 0$ et $d^{(1)}\geq 0$.
Rien n'est à prouver dans cette situation, les équations (F.8) et (F.9) coïncident.

2 : $d^{(1)}_{S_u\times T_u}\leq 0$ et $d^{(1)}\leq 0$.
Ce cas se réduit au premier cas grâce à la propriété suivante de $\mathrm{kl}(\cdot\|\cdot)$:

$$\mathrm{kl}\left(\frac{a+1}{2}\middle\|\frac{b+1}{2}\right)=\mathrm{kl}\left(\frac{-a+1}{2}\middle\|\frac{-b+1}{2}\right). \tag{F.10}$$

3 : $d^{(1)}_{S_u\times T_u}\leq 0$ et $d^{(1)}\geq 0$.
Dans cette situation, on a :

$$\begin{aligned}
&\mathrm{kl}\left(\frac{|d^{(1)}_{S_u\times T_u}|+1}{2}\middle\|\frac{|d^{(1)}|+1}{2}\right)-\mathrm{kl}\left(\frac{d^{(1)}_{S_u\times T_u}+1}{2}\middle\|\frac{d^{(1)}+1}{2}\right)\\
&\quad=\ \mathrm{kl}\left(\frac{-d^{(1)}_{S_u\times T_u}+1}{2}\middle\|\frac{d^{(1)}+1}{2}\right)-\mathrm{kl}\left(\frac{d^{(1)}_{S_u\times T_u}+1}{2}\middle\|\frac{d^{(1)}+1}{2}\right)\\
&\quad=\ \left(\frac{-d^{(1)}_{S_u\times T_u}+1}{2}-\frac{d^{(1)}_{S_u\times T_u}+1}{2}\right)\ln\left(\frac{1}{\frac{d^{(1)}+1}{2}}\right)\\
&\qquad+\left(\left(1-\frac{-d^{(1)}_{S_u\times T_u}+1}{2}\right)-\left(1-\frac{d^{(1)}_{S_u\times T_u}+1}{2}\right)\right)\ln\left(\frac{1}{1-\frac{d^{(1)}+1}{2}}\right)\\
&\quad=\ \left(-d^{(1)}_{S_u\times T_u}\right)\ln\left(\frac{1}{\frac{d^{(1)}+1}{2}}\right)\ +\ \left(d^{(1)}_{S_u\times T_u}\right)\ln\left(\frac{1}{1-\frac{d^{(1)}+1}{2}}\right)\\
&\quad=\ \left(-d^{(1)}_{S_u\times T_u}\right)\ln\left(\frac{1}{\frac{d^{(1)}+1}{2}}\right)\ +\ \left(d^{(1)}_{S_u\times T_u}\right)\ln\left(\frac{1}{\frac{-d^{(1)}+1}{2}}\right)\\
&\quad=\ d^{(1)}_{S_u\times T_u}\ \ln\left(\frac{d^{(1)}+1}{-d^{(1)}+1}\right)\\
&\quad\leq\ 0\,.
\end{aligned} \tag{F.11}$$

La dernière inégalité provient du fait que $d^{(1)}_{S_u \times T_u} \leq 0$ et $d^{(1)} \geq 0$.
Ainsi, des équations (F.11) et (F.8), on a :

$$\begin{aligned} \mathrm{kl}\left(\frac{|d^{(1)}_{S_u \times T_u}|+1}{2} \middle\| \frac{|d^{(1)}|+1}{2}\right) &\leq \mathrm{kl}\left(\frac{d^{(1)}_{S_u \times T_u}+1}{2} \middle\| \frac{d^{(1)}+1}{2}\right) \\ &\leq \frac{1}{m}\left[2\,\mathrm{KL}(\rho \,\|\, \pi) + \ln \frac{2\sqrt{m_u}}{\delta}\right], \end{aligned}$$

comme désiré.

4 : $d^{(1)}_{S_u \times T_u} \geq 0$ et $d^{(1)} \leq 0$.
L'équation (F.10) se réduit simplement au troisième cas.

F.3 Preuve du théorème 7.3

Nous proposons tout d'abord de majorer :

$$d^{(1)} = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{D_S}(h,h') - \mathbf{R}_{D_T}(h,h')\right],$$

par son estimation empirique :

$$d^{(1)}_{S\times T} = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathbf{R}_{S_u}(h,h') - \mathbf{R}_{T_u}(h,h')\right],$$

ainsi qu'un terme lié à la KL-divergence entre le posterior et le prior.

Pour ce faire, nous considérons le classifieur "abstrait" $\widehat{h} = (h,h') \in \mathcal{H}^2$ choisi aléatoirementselon la distribution $\widehat{\rho}$, où $\widehat{\rho}(\widehat{h}) = \rho(h)\rho(h')$. Signalons qu'avec $\widehat{\pi}(\widehat{h}) = \pi(h)\pi(h')$, on obtient avec l'équation (F.1) : $\mathrm{KL}(\widehat{\rho}\|\widehat{\pi}) = 2\,\mathrm{KL}(\rho\|\pi)$.
Nous définissons la fonction de perte "abstraite" pour $\widehat{h}$ pour une paire d'exemples $(\mathbf{x}^s, \mathbf{x}^t) \sim (D_S \times D_T)$ par :

$$\ell_{d^{(1)}}(\widehat{h}(\mathbf{x}^s), \widehat{h}(\mathbf{x}^t)) = \frac{1 + \ell_{0\text{-}1}(h(\mathbf{x}^s), h'(\mathbf{x}^s)) - \ell_{0\text{-}1}(h(\mathbf{x}^t), h'(\mathbf{x}^t))}{2}.$$

Ainsi, le risque "abstrait" de $\widehat{h}$ sur la distribution jointe défini par :

$$R^{(1)}_{D_S\times D_T}(\widehat{h}) = \mathop{\mathbf{E}}_{\mathbf{x}^s\sim D_S} \mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T} \ell_{d^{(1)}}(\widehat{h}(\mathbf{x}^s), \widehat{h}(\mathbf{x}^t)),$$

et le risque du classifieur de Gibbs associé est :

$$R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}}) = \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{D_S\times D_T}(\widehat{h}).$$

Les valeurs empiriques de ces deux quantités sont :

$$R^{(1)}_{S_u\times T_u}(\widehat{h}) = \mathop{\mathbf{E}}_{(\mathbf{x}^s,\mathbf{x}^t)\sim S_u\times T_u} \ell_{d^{(1)}}(\widehat{h}(\mathbf{x}^s), \widehat{h}(\mathbf{x}^t))$$

et,

$$R^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}}) = \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{S_u\times T_u}(\widehat{h}).$$

Il est facile de montrer que :

$$d^{(1)} = 2R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}}) - 1, \tag{F.12}$$

$$d^{(1)}_{S_u\times T_u} = 2R^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}}) - 1. \tag{F.13}$$

Puisque $\ell_{d^{(1)}}(\cdot,\cdot)$ renvoie des valeurs dans $[0,1]$, on peut borner $R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}})$ suivant la preuve du corollaire 3.3 (avec $C = 2A$). Pour ce faire, nous définissons la fonction convexe :

$$\mathcal{F}(b) = -\ln[1-(1-e^{-2A})b], \tag{F.14}$$

et nous considérons la variable aléatoire non négative :

$$\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\left(\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h})) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right).$$

On applique l'inégalité de Markov (théorème A.4, annexe A). Pour tout $\delta\in(0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix aléatoire de $S_u\times T_u\sim(D_S\times D_T)^{m_u}$, on a :

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\left(\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h})) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right)\\
&\leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S_u\times T_u\sim(D_S\times D_T)^{m_u}}\ \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\left(\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h})) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right).
\end{aligned}$$

En prenant le logarithme de chaque côté de l'inéquation précédente et en transformant de l'espérance sur $\widehat{\pi}$ en espérance selon $\widehat{\rho}$, on obtient :

$$\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}}\frac{\widehat{\pi}(\widehat{h})}{\widehat{\rho}(\widehat{h})} \exp\left(m_u\left(\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h})) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right)\right]\\
&\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S_u\times T_u\sim(D_S\times D_T)^{m_u}}\ \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\left(\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h})) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right)\right]\\
&= \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h}))\right)\mathop{\mathbf{E}}_{S_u\times T_u\sim(D_S\times D_T)^{m_u}} \exp\left(-2m_uAR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right].
\end{aligned} \tag{F.15}$$

Étant donné un classifieur $\widehat{h}$, nous définissons la variable aléatoire $Z_{\widehat{h}}$ qui suit une loi binomiale de m_u épreuves avec une probabilité de succès $R^{(1)}_{D_S\times D_T}(\widehat{h})$ et noté $B\left(m_u, R^{(1)}_{D_S\times D_T}(\widehat{h})\right)$. Le lemme de Maurer (lemme A.1) implique :

$$\begin{aligned}
\mathop{\mathbf{E}}_{S_u\times T_u\sim(D_S\times D_T)^{m_u}} \exp\left(-2m_uAR^{(1)}_{S_u\times T_u}(\widehat{h})\right) &\leq \mathop{\mathbf{E}}_{Z_{\widehat{h}}\sim B\left(m_u,R^{(1)}_{D_S\times D_T}(\widehat{h})\right)} \exp\left(-2AZ_{\widehat{h}}\right)\\
&= \sum_{k=0}^{m_u}\ \mathop{\mathbf{Pr}}_{Z_{\widehat{h}}\sim B\left(m_u,R^{(1)}_{D_S\times D_T}(\widehat{h})\right)}\left(Z_{\widehat{h}}=k\right)e^{-2Ak}\\
&= \sum_{k=0}^{m_u}\binom{m_u}{k}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})\right)^k\left(1-R^{(1)}_{S_u\times T_u}(\widehat{h})\right)^{m_u-k}e^{-2Ak}\\
&= \sum_{k=0}^{m_u}\binom{m_u}{k}\left(R^{(1)}_{S_u\times T_u}(\widehat{h})e^{-2A}\right)^k\left(1-R^{(1)}_{S_u\times T_u}(\widehat{h})\right)^{m_u-k}\\
&= \left[R^{(1)}_{S_u\times T_u}(\widehat{h})e^{-2A}+\left(1-R^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right]^{m_u}.
\end{aligned}$$

Le choix de $\mathcal{F}(\cdot)$ implique :

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} \exp\left(m_u \mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h}))\right) \mathop{\mathbf{E}}_{S_u\times T_u\sim(D_S\times D_T)^{m_u}} \exp\left(-2m_u A R^{(1)}_{S_u\times T_u}(\widehat{h})\right)\\
&\leq \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} e^{m_u\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h}))}\left[R^{(1)}_{S_u\times T_u}(\widehat{h})e^{-2A} + \left(1 - R^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right]^{m_u}\\
&= \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\pi}} 1\\
&= 1\,.
\end{aligned}$$

On majore maintenant l'équation (F.15) :

$$\ln\left[\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} \frac{\widehat{\pi}(\widehat{h})}{\widehat{\rho}(\widehat{h})} \exp\left(m_u\left(\mathcal{F}(R^{(1)}_{D_S\times D_T}(\widehat{h})) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right)\right] \leq \ln\frac{1}{\delta}\,.$$

On insère le terme $\mathrm{KL}(\rho\|\pi)$ dans la partie gauche de l'inégalité précédente en appliquant deux fois l'ingéalité de Jensen (théorème A.5, annexe A), une fois sur la fonction logarithme, puis sur la fonction convexe $\mathcal{F}(\cdot)$:

$$\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} \frac{\widehat{\pi}(\widehat{h})}{\widehat{\rho}(\widehat{h})} \exp\left(m_u\left(\mathcal{F}\left(R^{(1)}_{D_S\times D_T}(\widehat{h})\right) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right)\right]\\
&= \ln\left[\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} \exp\left(m_u\left(\mathcal{F}\left(R^{(1)}_{D_S\times D_T}(\widehat{h})\right) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right)\right)\right] - 2\,\mathrm{KL}(\rho\|\pi)\\
&\geq \mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} m\left(\mathcal{F}\left(R^{(1)}_{D_S\times D_T}(\widehat{h})\right) - 2AR^{(1)}_{S_u\times T_u}(\widehat{h})\right) - 2\,\mathrm{KL}(\rho\|\pi)\\
&\geq m_u\mathcal{F}\left(\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{D_S\times D_T}(\widehat{h})\right) - 2m_uA\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{S_u\times T_u}(\widehat{h}) - 2\,\mathrm{KL}(\rho\|\pi)\\
&= m_u\mathcal{F}\left(R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}})\right) - 2m_uAR^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}}) - 2\,\mathrm{KL}(\rho\|\pi)\,.
\end{aligned}$$

On a donc :

$$m_u\mathcal{F}\left(\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{D_S\times D_T}(\widehat{h})\right) - 2m_uA\mathop{\mathbf{E}}_{\widehat{h}\sim\widehat{\rho}} R^{(1)}_{S_u\times T_u}(\widehat{h}) - 2\,\mathrm{KL}(\rho\|\pi) \;\leq\; \ln\frac{1}{\delta}.$$

Ce qui implique :

$$\mathcal{F}(R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}})) \leq 2AR^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}}) + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{m_u}\,.$$

Maintenant en isolant le terme $R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}})$, on obtient :

$$R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}}) \leq \frac{1}{1-e^{-2A}}\left[1 - e^{-\left(2AR^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}}) + \frac{2\,\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m_u}\right)}\right],$$

l'inégalité $1 - e^{-x} \leq x$ implique :

$$R^{(1)}_{D_S\times D_T}(G_{\widehat{\rho}}) \leq \frac{1}{1-e^{-2A}}\left[2AR^{(1)}_{S_u\times T_u}(G_{\widehat{\rho}}) + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{m_u}\right].$$

D'après les équations (F.12) et (F.13), avec une probabilité d'au moins $1 - \frac{\delta}{2}$ sur le choix de $S \times T \sim (D_S \times D_T)^{m_u}$, on a :

$$\frac{d^{(1)} + 1}{2} \leq \frac{2A}{1 - e^{-2A}} \left[\frac{d^{(1)}_{S_u \times T_u} + 1}{2} + \frac{2\,\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{m \times 2A} \right],$$

De la même manière, on borne $d^{(2)} = \mathop{\mathbf{E}}_{(h,h') \sim \rho^2} [\mathbf{R}_{D_T}(h, h') - \mathbf{R}_{D_S}(h, h')]$. On obtient, avec une probabilité d'au moins $1 - \frac{\delta}{2}$ sur le choix $S \times T \sim (D_S \times D_T)^{m_u}$:

$$\frac{d^{(2)} + 1}{2} \leq \frac{2A}{1 - e^{-2A}} \left[\frac{d^{(2)}_{S_u \times T_u} + 1}{2} + \frac{2\,\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{m \times 2A} \right].$$

Pour finaliser la preuve, par définition, on a que $d^{(1)} = -d^{(2)}$, donc :

$$|d^{(1)}| = |d^{(2)}| = \mathrm{dis}_\rho(D_S, D_T),$$

et :

$$|d^{(1)}_{S_u \times T_u}| = |d^{(2)}_{S_u \times T_u}| = \mathrm{dis}_\rho(S, T).$$

Alors, la borne pour $\mathrm{dis}_\rho(D_S, D_T)$ est obtenue en prenant le maximum de la borne sur $d^{(1)}$ et de la borne sur $d^{(2)}$. Finalement, en appliquant la borne de l'union, avec une probabilité d'au moins $1 - \delta$ sur le choix de $S \times T \sim (D_S \times D_T)^{m_u}$, on a :

$$\frac{|d^{(1)}| + 1}{2} \leq \frac{A}{1 - e^{-2A}} \left[|d^{(1)}_{S_u \times T_u}| + 1 + \frac{2\,\mathrm{KL}(\rho \| \pi) + \ln \frac{2}{\delta}}{m_u A} \right],$$

où de manière équivalente :

$$\mathrm{dis}_\rho(D_S, D_T) \leq \frac{2A \left[\mathrm{dis}_\rho(S_u, T_u) + \frac{2\,\mathrm{KL}(\rho \| \pi) + \ln \frac{2}{\delta}}{m_u A} + 1 \right] - 1}{1 - e^{-2A}}.$$

Participations à la campagne d'évaluation TrecVid

G

Au nom du projet VideoSense nous avons participé à la tâche d'indexation sémantique dans le contexte des campages d'évalutations TrecVid [1] 2011 et 2012 [Smeaton *et al.*, 2006]. Vous trouvez dans les pages qui suivent le résumé de notre participation de l'année 2011.

1. http://trecvid.nist.gov/

VideoSense at TRECVID 2011 : Semantic Indexing from Light Similarity Functions-based Domain Adaptation with Stacking

Emilie Morvant[1], Stéphane Ayache[1], Amaury Habrard[2], Miriam Redi[3], Claudiu Tanase[3], Bernard Merialdo[3], Bahjat Safadi[4], Franck Thollard[4], Nadia Derbas[4], Georges Quenot[4]
[1] Aix-Marseille Univ, LIF-QARMA, CNRS UMR 7279, F-13013, Marseille, France
[2] University of St-Etienne, Lab. Hubert Curien, CNRS UMR 5516, F-42000, St-Etienne, France
[3] EURECOM, Sophia Antipolis, 2229 route des cretes, Sophia-Antipolis, France
[4] CNRS, LIG UMR 5217, Grenoble, F-38041, France

March 21, 2012

Abstract

This paper describes our participation to the TRECVID 2011 challenge [1]. This year, we focused on a stacking fusion with Domain Adaptation algorithm. In machine learning, Domain Adaptation deals with learning tasks where the train and the test distributions are supposed related but different. We have implemented a classical approach for concept detection using individual features (low-level and intermediate features) and supervised classifiers. Then we combine the various classifiers with a second layer of classifier (stacking) which was specifically designed for Domain Adaptation. We show that, empirically, Domain Adaptation can improve concept detection by considering test information during the learning process.

1 Introduction

The High-Level semantic retrieval task concerns features or concepts such as "Indoor/Outdoor", "People", "Speech" etc., that occur in video databases. The TRECVID SIN task [1] contributes to work on a benchmark for evaluating the effectiveness of detection methods for semantic concepts. The task is as follows: given the feature test collection composed of hundred of hours of videos, the common shot boundary reference for the feature extraction test collection, and the list of feature definitions, participants return for each feature the list of at most 2000 shots from the test collection, ranked according to the highest possibility of detecting the presence of the feature. Each feature is assumed to be binary, *i.e.*, it is either present or absent in the given reference shot.

The VideoSense[1] project aims at automatic video tagging by high level concepts, including static concepts (*e.g.* object, scene, people, etc.), events, and emotions, while targeting two applications, namely video recommendation and ads monetization. The innovations targeted by the project include video content description by low-level features, emotional video content recognition, cross-concept detection and multimodal fusion, and the use of a pivot language for dealing with multilingual textual resources associated with video data.

[1]The french project VideoSense ANR-09-CORD-026 of the ANR and the IST Programme of the European Community.

The first participation of the project to the TRECVID'11 SIN task is based on a stacking fusion with Domain Adaptation. Domain adaptation is a machine learning task consisting in adapting a classifier on new data that come from a distribution different but related to the distribution of the train examples. Since the TRECVID'11 corpus is a real corpus with amount data, the train data could not be representative of all the test data. Our aim is to test a Domain Adaptation algorithm that combines various classifier outputs from individual feature in order to adapt the learned classifier on the test data for improving the performances.

In the following, section 2 lists the features and classifiers we used. Section 3 presents our algorithm for Domain Adaptation namely DASF. Section 4 describes our fusion approach with DASF. Then, in section 5, we show the runs we submitted and discuss about their relative performance.

2 Feature extraction and individual classifiers

Since the VideoSense members are also members of the IRIM project which participate to TRECVID'11, features and individual classifiers are described in the paper [2].

We used the following features from LIG, EURECOM and LIF teams:

Global features: LIG/hg104 (Color histogram + Gabor filters)

Local features: LIG/opp_sift, LIG/stip (SIFT and STIP)

Shape features: EUR/sm462 (Saliency Moments)

Intermediate features: LIF/percepts (Mid-level concepts based on 15 concepts).

We used two types of classifiers from LIF team:

KNN: LIG_KNNB (multiclass KNN)

SVM: LIG_MSVM (for imbalanced data).

3 DASF: Domain Adaptation with Similarity Functions

Domain Adaptation arises when learning and test data are generated according to two different probability distributions: the first one generating training data is often referred to as the *source domain*, while the second one for test data corresponds to the *target domain*. According to the existing theoretical frameworks of Domain Adaptation [3, 4, 5] a classifier can perform well on the target domain if its error relatively to the source distribution and the divergence (or distance) between the source and target distributions are together low. One possible solution to learn a performing classifier on the target domain is to find a projection space in which the source and target distributions are close while keeping a low error on the source domain.

In order to illustrate quickly this idea in a formal manner, if err_T denotes the error on the target domain (ie our goal here) and err_S the error on the source domain where labeled data are available, then for any classifier h belonging to an hypothesis space $\mathcal{H}$, we have from [3]:

$$\forall h \in \mathcal{H},\ err_T(h) \ \leq \ err_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu. \tag{1}$$

The term ν can be seen as a kind of adaptation ability measure of $\mathcal{H}$ for the Domain Adaptation problem considered and corresponds to the error of the best joint hypothesis over the two domains: $\nu = \arg\min_{h\in\mathcal{H}} err_S(h) + err_T(h)$. The second term $d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$ is called the distance between the two domain marginal probability distributions: D_S for the source and D_T for the target. Note that both $err_S(h)$ and $d_{\mathcal{H}\Delta\mathcal{H}}$ can be estimated from finite samples.

Our algorithm addresses Domain Adaptation for binary classification in the challenging case where no target label is available. Following the Domain Adaptation framework of Ben-David *et al.* [3], our method looks for a relevant projection space where the source and target distributions tend to be close. Our approach is formulated as a linear program with a 1-norm regularization leading to sparse models. To improve the efficiency of the method we propose an iterative version based on a reweighting scheme of the similarities to move closer the distributions in a new projection space. Hyperparameters and reweighting quality are controlled by a reverse validation process.

Furthermore, our approach is based on a recent framework of Balcan *et al.* [6] allowing to learn linear classifiers in an explicit projection space based on *good similarity functions* defined as follows. Roughly speaking, under a criterion of goodness introduced in [6], a good similarity function ensures that a low error linear classifier exists in a space made of similarity scores to a set of prototype examples called reasonable points. The learned linear classifier is of the form:

$$h(\cdot) = \sum_{i=1}^{d_u} \alpha_i K(\cdot, \mathbf{x}'_i)$$

where the examples $\mathbf{x}_i$ correspond to this so called set of reasonable points. The formulation is close to the one of SVM, except that the framework allows one to use similarity functions K that are more general than kernels in the sense K is not required to be symmetric nor positive semi-definite. The main idea of our Domain Adaptation algorithm, called DASF, consists in automatically modifying the projection space to the similarities (ie $\phi^R(\cdot) =< K(\cdot, \mathbf{x}'_1), \ldots, K(\cdot, \mathbf{x}'_i), \ldots, K(\cdot, \mathbf{x}'_{d_u}) >$) for moving closer source and target points. For this purpose, we proposed a general method based on the optimization of a regularized convex objective function trying to find reasonable points close both to source and target examples. The objective function proposed in our algorithm is defined as follows:

$$\left\{ \begin{array}{rcl} \min\limits_{\boldsymbol{\alpha}} \; F(\boldsymbol{\alpha}) & = & \dfrac{1}{d_l} \displaystyle\sum_{i=1}^{d_l} L\big(h, (\mathbf{x}_i, y_i)\big) + \lambda \|\boldsymbol{\alpha}\|_1 \\ & & + \beta \displaystyle\sum_{(\mathbf{x}_s, \mathbf{x}_t)\in\mathcal{C}_{ST}} \Big\| \big({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t)\big) \operatorname{diag}(\boldsymbol{\alpha}) \Big\|_1, \\ \text{with } L\big(h, (\mathbf{x}_i, y_i)\big) = \Big[1 - y_i h(\mathbf{x}_i)\Big]_+ & \text{and} & h(\mathbf{x}_i) = \displaystyle\sum_{j=1}^{d_u} \alpha_j K(\mathbf{x}_i, \mathbf{x}'_j). \end{array} \right. \qquad (DASF_{opt})$$

Examples $(\mathbf{x}_i, y_i)$ are labeled examples from the source domain, $L(\cdot)$ is the loss function optimized for learning classifiers, $\mathbf{x}'_j$ being the "reasonable" points considered. We have two regularizers, one is a classical $L1$ norm over the weights α_j of the linear classifier $\boldsymbol{\alpha}$. The complex last one, weighted by a parameter β, corresponds to the term trying to move closer some source instances $(\mathbf{x}_s)$ with target examples $(\mathbf{x}_t)$ belonging to a set chosen by reverse validation.

More details can be found in [7]. Note that we have provided theoretical results with DASF and that our algorithm has been evaluated on a toy problem and on two real image annotation tasks.

4 Classifier fusion with DASF

We used DASF algorithm as a stacking classifier using outputs from individual classifier on the considered features. Stacking provides a way of combining classifiers together to find an overall system with usually improved generalization performance [8]. In the context of SIN task, we considered the training set as source domain and the test set as target domain, although they are supposed to be already closed. However, as the challenge is based on real data, the training set could not be representative of the test set. We thus expect DASF to improve further the performance of video indexing.

In order to make good use of the similarity function framework of Balcan *et al.* we proposed to compare two different similarity function. The first one is a usual Gaussian kernel (which is symmetric and PSD). For the second one we build a new similarity function by normalizing the Gaussian kernel. In order to link the two domains, we consider the information of both of them at the same time by actually renormalizing the Gaussian kernel for all the source and target points relatively to the set of similarities to the reasonable points. By construction, the similarity is then non-symmetric and non-PSD. Our choice is clearly heuristic and our aim is just to evaluate the interest of renormalizing a similarity for domain adaptation problems. Finally, we made a light search of the hyperparameters to accelerate our approach.

5 Results

In the following, we report our results (infAP) on the full SIN task. We compare DASF with SF as our baseline, SF-classifier is a linear classifier without Domain Adaptation, proposed for leaning with good similarity function ([6]). Both SF and DASF have been submitted with and without normalization. We can observe that:

- A poor performance of our runs compare to the best run and the median. The stacking process as we implemented seems to overfit data ;
- Normalization was not successful, probably because source and target domains was actually closed: in such a particular case, we lost information by making use of the normalization ;
- As expected Domain Adaptation runs both outperformed our baseline. Even if train and test set are similar, we can still take advantage of Domain Adaptation approach by considering (test) target information during the learning process.

Videosense 1	0.067	DASF with normalization
Videosense 2	0.040	SF with normalization
Videosense 3	0.080	DASF
Videosense 4	0.072	SF
Best run	0.1731	-
Median	0.1083	-

6 Conclusion

Our main focus for TRECVID'11 SIN task is on the exploration of Stacking with Domain Adaptation to combine individual classifiers. Taking into account target information by using Domain Adaptation has allowed us to improve baseline results which was our main objective. However, we suspect that we actually have overfitted data and unfortunately was not able to generalize as we expected. This can be explained by the following reasons:

- We did not consider the goodness of the similarity function used (ie the Gaussian kernel) according to the outputs of the individual classifiers. In particular we are not sure that such a similarity is the best choice with our framework. Optimizing the similarity with some metric learning approaches, for example, would help us to better adapt our framework to the stacking problem considered.
- The renormalization of the kernel function used in our approach is simple but we know that it works well when the two domains are very different. The fact that this renormalization does not increase the performance shows that the train and test data are not that different here. By combining an information on the distance domains and some metric learning approaches, as evoked before, we may drastically improve the results.
- In domain adaptation, the estimation of the hyperparameters is difficult and costly. In our runs, we simplified a bit the search of the parameters which may explain why we tend to overfit.
- We did not take into account some second order (ie variance/covariance) information from the outputs of the classifiers and we think that it may be useful to find a better combination.

Our first attempt shows that Domain Adaptation could bring some improvement. However, to be competitive with the state of the art, we have to take into account more accurately the diversity of the data and classifiers. It appears especially important in the case of multiple source of annotations. Under this condition, we think that Domain Adaptation approaches can lead to significant improvement of the results. Using some correlations between labels is in particular an interesting direction.

Acknowledgment

This work was supported in part by the french project VideoSense ANR-09-CORD-026 of the ANR.

References

[1] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[2] Bertrand Delezoide, Frédéric Precioso, Philippe Gosselin, Miriam Redi, Bernard Mérialdo, Lionel Granjon, Denis Pellerin, Michèle Rombaut, Hervé Jégou, Rémi Vieux, Boris Mansencal, Jenny Benois-Pineau, Stéphane Ayache, Bahjat Safadi, Franck Thollard, Georges Quénot, Hervé Bredin, Matthieu Cord, Alexandre Benoit, Patrick Lambert, Tiberius Strat, Joseph Razik, Sébastion Paris, and Hervé Glotin. Irim at trecvid 2011: Semantic indexing and instance search. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Machine Learning Journal*, 79(1-2):151–175, 2010.

[4] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of COLT*, pages 19–30, 2009.

[5] S. Ben-David, T. Lu, T. Luu, and D. Pal. Impossibility theorems for domain adaptation. *JMLR W&CP*, 9:129–136, 2010.

[6] M.-F. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. In *Proceedings of COLT*, pages 287–298, 2008.

[7] Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Sparse Domain Adaptation in Projection Spaces based on Good Similarity Functions. In *11th IEEE International Conference on Data Mining (ICDM)*, pages 457–466. IEEE Computer Society, 2011.

[8] Padhraic Smyth and David Wolpert. Linearly combining density estimators via stacking. *Machine Learning*, 36(1-2):59–83, 1999.

Bibliographie

[Ambroladze *et al.*, 2006] A. Ambroladze, E. Parrado-Hernández, et J. Shawe-Taylor. Tighter PAC-Bayes bounds. Dans *Proceedings of Annual Conference on Neural Information Processing Systems*, pages 9–16, 2006. (Cité pages 58, 62, 65, 100 et 164.)

[Atrey *et al.*, 2010] P. K. Atrey, M. Anwar Hossain, A. El-Saddik, et M. S. Kankanhalli. Multimodal fusion for multimedia analysis : a survey. *Multimedia System*, 16(6) :345–379, 2010. (Cité pages 91, 92 et 93.)

[Ayache *et al.*, 2007] S. Ayache, G. Quénot, et J. Gensel. Image and video indexing using networks of operators. *Journal on Image and Video Processing*, 2007 :1 :1–1 :13, 2007. (Cité page 147.)

[Balcan *et al.*, 2004] M.-F. Balcan, A. Blum, et Y. Ke. Co-training and expansion : Towards bridging theory and practice. *Computer Science Department*, page 154, 2004. (Cité page 54.)

[Balcan *et al.*, 2008a] M. F. Balcan, A. Blum, et N. Srebro. Improved guarantees for learning via similarity functions. Dans *Proceedings of Annual Conference on Computational Learning Theory*, pages 287–298, 2008. (Cité pages 5, 24, 31, 33, 34, 126, 127 et 153.)

[Balcan *et al.*, 2008b] M. F. Balcan, A. Blum, et N. Srebro. A theory of learning with similarity functions. *Machine Learning Journal*, 72(1-2) :89–112, 2008. (Cité pages 5, 24, 31, 126, 127 et 153.)

[Banerjee, 2006] A. Banerjee. On bayesian bounds. Dans *Proceedings of International Conference on Machine Learning*, pages 81–88, 2006. (Cité page 63.)

[Bardenet *et al.*, 2013] R. Bardenet, M. Brendel, B. Kégl, et M. Sebag. Collaborative hyperparameter tuning. Dans *Proceedings of International Conference on Machine Learning*, 2013. (Cité page 177.)

[Bartlett et Mendelson, 2002] P. L. Bartlett et S. Mendelson. Rademacher and gaussian complexities : Risk bounds and structural results. *Journal of Machine Learning Research*, 3 :463–482, 2002. (Cité pages 20 et 21.)

[Bellet *et al.*, 2013a] A. Bellet, A. Habrard, E. Morvant, et M. Sebban. Vote de majorité a priori contraint pour la classification binaire : spécification au cas des plus proches voisins. Dans *Conférence Francophone sur l'Apprentissage Automatique*, 2013. (Cité page 76.)

[Bellet *et al.*, 2011] A. Bellet, A. Habrard, et M. Sebban. Learning good edit similarities with generalization guarantees. Dans *Proceedings of European Conference on Machine Learning and Principles of Data Mining and Knowledge Discovery*, volume 6911 de *LNCS*, pages 188–203, 2011. (Cité page 176.)

[Bellet *et al.*, 2012] A. Bellet, A. Habrard, et M. Sebban. Similarity learning for provably accurate sparse linear classification. Dans *Proceedings of International Conference on Machine Learning*, 2012. (Cité page 89.)

[Bellet *et al.*, 2013b] A. Bellet, A. Habrard, et M. Sebban. A survey on metric learning for feature vectors and structured data. *ArXiv e-prints*, 2013. `http://arxiv.org/abs/1306.6709`. (Cité page 27.)

[Ben-David *et al.*, 2010] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, et J.W. Vaughan. A theory of learning from different domains. *Machine Learning Journal*, 79(1-2) :151–175, 2010. (Cité pages 4, 45, 46, 47, 48, 49, 50, 137, 139, 159, 161 et 175.)

[Ben-David *et al.*, 2007] S. Ben-David, J. Blitzer, K. Crammer, et F. Pereira. Analysis of representations for domain adaptation. Dans *Proceedings of Annual Conference on Neural Information Processing Systems*, pages 137–144, 2007. (Cité pages 4, 45, 46, 47, 48, 49 et 159.)

[Ben-David *et al.*, 2012a] S. Ben-David, D. Loker, N. Srebro, et K. Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. Dans *Proceedings of International Conference on Machine Learning*, 2012. (Cité page 16.)

[Ben-David *et al.*, 2010] S. Ben-David, T. Lu, T. Luu, et D. Pal. Impossibility theorems for domain adaptation. *JMLR W&CP, Proceedings of International Conference on Artificial Intelligence and Statistics*, 9 :129–136, 2010. (Cité pages 41, 42, 44 et 51.)

[Ben-David *et al.*, 2012b] S. Ben-David, S. Shalev-Shwartz, et R. Urner. Domain adaptation–can quantity compensate for quality ? Dans *Proceedings of International Symposium on Artificial Intelligence and Mathematics*, 2012. (Cité pages 42 et 43.)

[Ben-David et Urner, 2012] S. Ben-David et R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. Dans *Proceedings of Algorithmic Learning Theory*, pages 139–153, 2012. (Cité pages 41 et 44.)

[Bengio, 2009] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1) :1–127, 2009. (Cité page 2.)

[Bishop *et al.*, 2006] C. M. Bishop et others. *Pattern recognition and machine learning*, volume 1. springer New York, 2006. (Cité page 1.)

[Blanchard et Fleuret, 2007] G. Blanchard et F. Fleuret. Occam's hammer. Dans *Proceedings of Annual Conference on Learning Theory*, pages 112–126, 2007. (Cité pages 58 et 171.)

[Blitzer *et al.*, 2007] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, et J. Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20 :129–136, 2007. (Cité page 175.)

[Blitzer *et al.*, 2011] J. Blitzer, D.n Foster, et S. Kakade. Domain adaptation with coupled subspaces. *Journal of Machine Learning Research-Proceedings Track*, 15 :173–181, 2011. (Cité page 40.)

[Blitzer *et al.*, 2006] J. Blitzer, R. McDonald, et F. Pereira. Domain adaptation with structural correspondence learning. Dans *Proceedings of Conference on Empirical Methods on Natural Language Processing*, pages 120–128, 2006. (Cité pages 40 et 168.)

[Boser *et al.*, 1992] B. E. Boser, I. M. Guyon, et V. N. Vapnik. A training algorithm for optimal margin classifiers. Dans *Proceedings of the annual workshop on Computational learning theory*, pages 144–152, 1992. (Cité pages 3 et 27.)

[Boucheron *et al.*, 2004] S. Boucheron, G. Lugosi, et O. Bousquet. Concentration inequalities. Dans *Advanced Lectures on Machine Learning*, volume 3176 de *Lecture Notes in Computer Science*, pages 208–240, 2004. (Cité page 19.)

[Bousquet et Elisseeff, 2002] O. Bousquet et A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2 :499–526, 2002. (Cité page 22.)

[Bregman, 1967] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3) :200–217, 1967. (Cité page 45.)

[Breiman, 2001] L. Breiman. Random Forests. *Machine Learning*, 45(1) :5–32, October 2001. (Cité page 115.)

[Breiman *et al.*, 1984] L. Breiman, J. Friedman, R. Olshen, , et C. Stone. Classification and regression trees. *Wadsworth International Group*, 1984. (Cité page 2.)

[Bruzzone et Marconcini, 2010] L. Bruzzone et M. Marconcini. Domain adaptation problems : A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5) :770–787, 2010. (Cité pages 40, 53, 55, 132, 140 et 167.)

[C. Zhang, 2012] J. Ye C. Zhang, L. Zhang. Generalization bounds for domain adaptation. Dans *Proceedings of Annual Conference on Neural Information Processing Systems*, 2012. (Cité pages 45 et 175.)

[Cao *et al.*, 2011] B. Cao, X. Ni, J.-T. Sun, G. Wang, et Q. Yang. Distance metric learning under covariate shift. Dans *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1204–1210, 2011. (Cité page 176.)

[Catoni, 2003] O. Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, (840), 2003. (Cité page 62.)

[Catoni, 2007] O. Catoni. *PAC-Bayesian supervised classification : the thermodynamics of statistical learning*, volume 56. Institute of Mathematical Statistic, 2007. (Cité pages 61 et 64.)

[Cesa-Bianchi *et al.*, 2004] N. Cesa-Bianchi, A. Conconi, et C. Gentile. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50(9) :2050–2057, 2004. (Cité page 176.)

[Chang et Lin, 2001] C.-C. Chang et C.-J. Lin. *LIBSVM : a library for support vector machines*, 2001. (Cité pages 97, 140 et 167.)

[Chen *et al.*, 2011a] M. Chen, K. Q. Weinberger, et J. Blitzer. Co-training for domain adaptation. Dans *Proceedings of Annual Conference on Neural Information Processing Systems*, pages 2456–2464, 2011. (Cité pages 41, 54, 55, 167 et 169.)

[Chen *et al.*, 2011b] M. Chen, K. Q. Weinberger, et Y. Chen. Automatic feature decomposition for single view co-training. Dans *Proceedings of International Conference on Machine Learning*, 2011. (Cité page 54.)

[Cornuéjols et Miclet, 2010] A. Cornuéjols et L. Miclet. *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles, 2010. (Cité pages 1 et 25.)

[Cortes et Mohri, 2011] C. Cortes et M. Mohri. Domain adaptation in regression. Dans *Proceedings of Algorithmic Learning Theory*, pages 308–323, 2011. (Cité page 47.)

[Cortes et Vapnik, 1995] C. Cortes et V. Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995. (Cité page 27.)

[Cover et Hart, 1967] T. Cover et P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1) :21–27, 1967. (Cité pages 2 et 25.)

[Crammer *et al.*, 2008] K. Crammer, M. Kearns, et J. Wortman. Learning from multiple sources. *The Journal of Machine Learning Research*, 9 :1757–1774, 2008. (Cité pages 37 et 174.)

[Dai *et al.*, 2007] W. Dai, Q. Yang, G.-R. Xue, et Y. Yu. Boosting for transfer learning. Dans *Proceedings of International Conference on Machine learning*, pages 193–200, 2007. (Cité page 40.)

[Daumé III, 2007] H. Daumé III. Frustratingly easy domain adaptation. Dans *ACL*, 2007. (Cité pages 36, 37 et 41.)

[Daumé III *et al.*, 2010] H. Daumé III, A. Kumar, et A. Saha. Co-regularization based semi-supervised domain adaptation. Dans *NIPS*, pages 478–486, 2010. (Cité pages 36 et 41.)

[Devroye *et al.*, 1996] L. Devroye, L. Györfi, et G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. (Cité page 85.)

[Dietterich, 2000] T. G. Dietterich. Ensemble methods in machine learning. Dans *Multiple Classifier Systems*, pages 1–15, 2000. (Cité pages 92 et 93.)

[Donsker et Varadhan, 1975] D. Donsker et S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975. (Cité page 111.)

[Duan *et al.*, 2009] Lixin Duan, Ivor W Tsang, Dong Xu, et Stephen J Maybank. Domain transfer svm for video concept detection. Dans *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 1375–1381, 2009. (Cité page 36.)

[Duda *et al.*, 2001] R.O. Duda, P.E. Hart, et D.G. Stork. *Pattern classification*. Pattern Classification and Scene Analysis : Pattern Classification. Wiley, 2001. (Cité pages 25 et 85.)

[Everingham *et al.*, 2007] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, et A. Zisserman. The PASCAL visual object classes challenge 2007 (VOC2007) results. 2007. (Cité pages 96 et 146.)

[Evgeniou *et al.*, 2006] T. Evgeniou, C. A. Micchelli, et M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(1) :615, 2006. (Cité page 176.)

[Floyd et Warmuth, 1995] S. Floyd et M. K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning Journal*, 21(3) :269–304, 1995. (Cité pages 72, 80 et 81.)

[Freund et Schapire, 1996] Y. Freund et R. E. Schapire. Experiments with a new boosting algorithm. Dans *Proceedings of International Conference on Machine Learning*, pages 148–156, 1996. (Cité pages 2 et 92.)

[Fürnkranz et Hüllermeier (eds), 2010] J. Fürnkranz et E. Hüllermeier (eds). *Preference Learning*. Springer-Verlag, 2010. (Cité page 94.)

[Gelman *et al.*, 2004] A. Gelman, J. B. Carlin, H. S. Stern, et D. B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 2004. (Cité page 58.)

[Geng *et al.*, 2011] B. Geng, D. Tao, et C. Xu. DAML : Domain adaptation metric learning. *IEEE Transactions on Image Processing*, 20(10) :2980–2989, 2011. (Cité page 176.)

[Geras et Sutton, 2013] K. J. Geras et C. Sutton. Multiple-source cross-validation. Dans *Proceedings of International Conference on Machine Learning*, 2013. (Cité page 55.)

[Germain *et al.*, 2013a] P. Germain, A. Habrard, F. Laviolette, et E. Morvant. PAC-bayesian domain adaptation bound with specialization to linear classifiers. Dans *Proceedings of International Conference on Machine Learning*, 2013. (Cité page 156.)

[Germain *et al.*, 2013b] P. Germain, A. Habrard, F. Laviolette, et E. Morvant. Une analyse PAC-Bayésienne de l'adaptation de domaine et sa spécialisation aux classifieurs linéaires. Dans *Conférence Francophone sur l'Apprentissage Automatique*, 2013. (Cité page 156.)

[Germain *et al.*, 2009a] P. Germain, A. Lacasse, F. Laviolette, et M. Marchand. PAC-Bayesian Learning of Linear Classifiers. Dans *Proceedings of International Conference on Machine Learning*, 2009. (Cité pages 59, 61, 62, 65, 66, 165 et 167.)

[Germain *et al.*, 2009b] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, et S. Shanian. From PAC-Bayes bounds to kl regularization. Dans *Proceedings of annual conference on Neural Information Processing Systems*, pages 603–610, 2009. (Cité page 62.)

[Germain *et al.*, 2011] P. Germain, A. Lacoste, F. Laviolette, M. Marchand, et S. Shanian. A PAC-Bayes sample compression approach to kernel methods. Dans *Proceedings of International Conference on Machine Learning*, 2011. (Cité pages 62 et 81.)

[Giguère *et al.*, 2013] S. Giguère, F. Laviolette, M. Marchand, et K. Sylla. Risk bounds and learning algorithms for the regression approach to structured output prediction. Dans *Proceedings of International Conference on Machine Learning*, 2013. (Cité pages 63 et 177.)

[Glorot *et al.*, 2011] X. Glorot, A. Bordes, et Y. Bengio. Domain adaptation for large-scale sentiment classification : A deep learning approach. Dans *Proceedings of International Conference on Machine Learning*, 2011. (Cité page 177.)

[Gong *et al.*, 2013] B. Gong, K. Grauman, et F. Sha. Connecting the dots with landmarks : Discriminatively learning domain-invariant features for unsupervised domain adaptation. Dans *Proceedings of International Conference on Machine Learning*, 2013. (Cité page 153.)

[Graepel *et al.*, 2005] T. Graepel, R. Herbrich, et J. Shawe-Taylor. PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning Journal*, 59(1-2) :55–76, 2005. (Cité pages 62, 72 et 81.)

[Guermeur, 2007] Y. Guermeur. *SVM Multiclasses, Théorie et Applications*. Habilitation à diriger des recherches, Université Nancy 1, 2007. (Cité page 27.)

[Habrard *et al.*, 2011] A. Habrard, J.-P. Peyrache, et M. Sebban. Domain adaptation with good edit similarities : A sparse way to deal with scaling and rotation problems in image classification. Dans *Proceedings of Conference on Tools with Artificial Intelligence*, pages 181–188. IEEE, 2011. (Cité page 53.)

[Habrard *et al.*, 2013] A. Habrard, J.-P. Peyrache, et M. Sebban. Boosting for unsupervised domain adaptation. Dans *Proceedings of European Conference on Machine Learning and Principles of Data Mining and Knowledge Discovery*, 2013. (Cité page 40.)

[Harel et Mannor, 2012] M. Harel et S. Mannor. The perturbed variation. Dans *NIPS*, pages 1943–1951, 2012. (Cité page 45.)

[Hastie et Tibshirani, 1996] T. Hastie et R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6) :607–616, juin 1996. (Cité pages 27 et 83.)

[Higgs et Shawe-Taylor, 2010] M. Higgs et J. Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. Dans *Proceedings of Algorithmic Learning Theory*, pages 148–162, 2010. (Cité page 63.)

[Hsu *et al.*, 2012] D. Hsu, S. Kakade, et T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17, 2012. (Cité pages 121 et 177.)

[Huang *et al.*, 2007] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, et B. Scholkopf. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19 :601, 2007. (Cité pages 39 et 45.)

[Jiang, 2008] J. Jiang. A literature survey on domain adaptation of statistical classifiers. Rapport Technique, CS Department at University of Illinois at Urbana-Champaign, 2008. (Cité page 35.)

[Jiang et Zhai, 2007] J. Jiang et C. Zhai. Instance weighting for domain adaptation in nlp. Dans *Proceedings of Association for Computational Linguistics*, 2007. (Cité page 39.)

[Jin et Wang, 2012] C. Jin et L. Wang. Dimensionality dependent PAC-Bayes margin bound. Dans *Advances in Neural Information Processing Systems 25*, pages 1043–1051, 2012. (Cité page 66.)

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. Dans *Proceedings of International Conference on Machine Learning*, pages 200–209, 1999. (Cité pages 140 et 167.)

[Kaelbling *et al.*, 1996] L. P. Kaelbling, M. L. Littman, et A. W. Moore. Reinforcement learning : A survey. *arXiv preprint cs/9605103*, 1996. (Cité page 178.)

[Kedem *et al.*, 2012] D. Kedem, S. Tyree, K. Weinberger, F. Sha, et G. Lanckriet. Nonlinear metric learning. Dans *Proceedings of Annual Conference on Neural Information Processing Systems*, volume 25, pages 2582–2590, 2012. (Cité page 100.)

[Kifer *et al.*, 2004] D. Kifer, S. Ben-David, et J. Gehrke. Detecting change in data streams. Dans *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 180–191, 2004. (Cité pages 45 et 46.)

[Kittler *et al.*, 1998] J. Kittler, M. Hatef, R. P. W. Duin, et J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 :226–239, 1998. (Cité page 92.)

[Kolmogorov et Tikhomirov, 1959] A. Nikolaevich Kolmogorov et V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2) :3–86, 1959. (Cité page 23.)

[Koltchinskii et Panchenko, 1999] V. Koltchinskii et D. Panchenko. Rademacher processes and bounding the risk of function learning. Dans *High Dimensional Probability II*, pages 443–459. Birkhäuser, 1999. (Cité pages 20 et 21.)

[Kubat *et al.*, 1997] M. Kubat, S. Matwin, et others. Addressing the curse of imbalanced training sets : one-sided selection. Dans *Machine Learning-International Workshop then Conference-*, pages 179–186. Morgan Kaufmann Publishers, inc., 1997. (Cité page 39.)

[Kulis *et al.*, 2011] B. Kulis, K. Saenko, et T. Darrell. What you saw is not what you get : Domain adaptation using asymmetric kernel transforms. Dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1785–1792, 2011. (Cité page 176.)

[Kullback et Leibler, 1951] S. Kullback et R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1) :79–86, 1951. (Cité page 45.)

[Kuncheva, 2004] L. I. Kuncheva. *Combining Pattern Classifiers : Methods and Algorithms*. Wiley-Interscience, 2004. (Cité pages 92, 93 et 94.)

[Lacasse, 2010] A. Lacasse. *Bornes PAC-Bayes et algorithmes d'apprentissage*. Thèse, Université Laval, 2010. (Cité page 65.)

[Lacasse *et al.*, 2007] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, et N. Usunier. PAC-bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. Dans *Proceedings of annual conference on Neural Information Processing Systems*, 2007. (Cité pages 60, 67 et 69.)

[Langford, 2005] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6 :273–306, 2005. (Cité pages 58, 61, 62, 63, 66, 106 et 110.)

[Langford et Shawe-Taylor, 2002] J. Langford et J. Shawe-Taylor. PAC-bayes & margins. Dans *Proceedings of Annual Conference on Advances in Neural Information Processing Systems*, pages 439–446. MIT Press, 2002. (Cité pages 58, 60, 65, 164 et 175.)

[Laviolette et Marchand, 2007] F. Laviolette et M. Marchand. PAC-Bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers. *Journal of Machine Learning Research*, 8 :1461–1487, 2007. (Cité pages 62, 72, 80 et 81.)

[Laviolette *et al.*, 2011a] F. Laviolette, M. Marchand, et J.-F. Roy. From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. Dans *Proceedings of International Conference on Machine Learning*, June 2011. (Cité pages 5, 59, 60, 62, 67, 68, 69, 70, 71, 72, 75, 77, 80 et 87.)

[Laviolette *et al.*, 2011b] F. Laviolette, M. Marchand, et J.-F. Roy. *From PAC-Bayes Bounds to Quadratic Programs for Majority Votes : Extended version*, 2011. (Cité pages 70 et 187.)

[Lettvin *et al.*, 1959] J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, et W. H. Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11) :1940–1951, 1959. (Cité page 2.)

[Lever *et al.*, 2010] G. Lever, F. Laviolette, et J. Shawe-Taylor. Distribution-dependent PAC-Bayes priors. Dans *Proceedings of Algorithmic Learning Theory*, pages 119–133, 2010. (Cité pages 62 et 175.)

[Lever *et al.*, 2013] G. Lever, F. Laviolette, et J. Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473 :4–28, 2013. (Cité pages 62 et 175.)

[Li et Bilmes, 2007] X. Li et J. Bilmes. A bayesian divergence prior for classifier adaptation. Dans *JMLR W&CP, Proceedings of International Conference on Artificial Intelligence and Statistics*, 2007. (Cité page 45.)

[Lin *et al.*, 2002] Y. Lin, Y. Lee, et G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46(1-3) :191–202, 2002. (Cité page 39.)

[Liu *et al.*, 2008] Q. Liu, A. J. Mackey, D. S. Roos, et F. C. N. Pereira. Evigan : a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics*, 24(5) :597–605, 2008. (Cité page 36.)

[Lowe, 1999] D. G. Lowe. Object recognition from local scale-invariant features. Dans *ICCV*, 1999. (Cité page 90.)

[Machart, 2012] P. Machart. *Coping with the Computational and Statistical Bipolar Nature of Machine Learning*. Thèse, Aix-Marseille Université, 2012. (Cité page 102.)

[Mairal, 2010] J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. thèse, École normale supérieure de Cachant, 2010. (Cité page 129.)

[Mansour *et al.*, 2009a] Y. Mansour, M. Mohri, et A. Rostamizadeh. Domain adaptation : Learning bounds and algorithms. Dans *Proceedings of Annual Conference on Learning Theory*, pages 19–30, 2009. (Cité pages 4, 45, 46, 47, 49, 159 et 161.)

[Mansour *et al.*, 2009b] Y. Mansour, M. Mohri, et A. Rostamizadeh. Multiple source adaptation and the rényi divergence. Dans *Proceedings of annual Conference on Uncertainty in Artificial Intelligence*, pages 367–374, 2009. (Cité pages 39, 45 et 175.)

[Mansour et Schain, 2012] Y. Mansour et M. Schain. Robust domain adaptation. Dans *Proceedings of International Symposium on Artificial Intelligence and Mathematics*, 2012. (Cité pages 43 et 44.)

[Marchand et Sokolova, 2005] M. Marchand et M. Sokolova. Learning with decision lists of data-dependent features. *Journal of Machine Learning Research*, 6 :427–451, 2005. (Cité page 81.)

[Margolis, 2011] A. Margolis. A literature review of domain adaptation with unlabeled data. Rapport Technique, University of Washington, 2011. (Cité page 35.)

[Maurer, 2004] Andreas Maurer. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004. (Cité page 181.)

[McAllester, 1999] D. A. McAllester. Some PAC-bayesian theorems. *Machine Learning Journal*, 37 :355–363, 1999. (Cité pages 4, 57, 61 et 64.)

[McAllester, 2003] D. A. McAllester. Simplified PAC-bayesian margin bounds. Dans *Proceedings of annual conference on Computational learning theory*, pages 203–215, 2003. (Cité pages 61, 64, 106, 110, 111, 187 et 190.)

[McClosky *et al.*, 2006] D. McClosky, E. Charniak, et M. Johnson. Reranking and self-training for parser adaptation. Dans *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 337–344. Association for Computational Linguistics, 2006. (Cité page 36.)

[Micchelli et Pontil, 2005] C. A. Micchelli et M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1) :177–204, 2005. (Cité pages 29 et 177.)

[Mitchell, 1982] T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2) :203–226, 1982. (Cité page 13.)

[Mitchell, 1997] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 édition, 1997. (Cité page 1.)

[Mohri et Medina, 2012] M. Mohri et A. M. Medina. New analysis and algorithm for learning with drifting distributions. Dans *Algorithmic Learning Theory*, pages 124–138. Springer, 2012. (Cité page 175.)

[Mohri *et al.*, 2012] M. Mohri, A. Rostamizadeh, et A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. (Cité page 1.)

[Morvant *et al.*, 2011a] E. Morvant, S. Ayache, et A. Habrard. Adaptation de domaine parcimonieuse par pondération de bonnes fonctions de similarité. Dans Presses de L'université des Antilles et de la Guyanne, éditeurs, *Conférence Francophone d'Apprentissage*, Sciences exactes et naturelles, pages 295–310. Publibook, 2011. (Cité page 126.)

[Morvant *et al.*, 2011b] E. Morvant, A. Habrard, et S. Ayache. On the usefulness of similarity based projection spaces for transfer learning. Dans *Proceedings of the 1st Similarity-Based Patterns Recognition workshop*, volume 7005 de *LNCS*, pages 1–16. Springer, 2011. (Full Paper, Acceptance rate : 32%). (Cité page 126.)

[Morvant *et al.*, 2011c] E. Morvant, A. Habrard, et S. Ayache. Sparse domain adaptation in projection spaces based on good similarity functions. Dans *Proceedings of the 11th IEEE International Conference on Data Mining series*, pages 457–466. IEEE Computer Society, 2011. (Full Paper, Acceptance Rate : 18%) Selected as one of the best papers for possible publication in Knowledge and Information Systems. (Cité page 126.)

[Morvant *et al.*, 2012a] E. Morvant, A. Habrard, et S. Ayache. Étude de la généralisation de DASF à l'adaptation de domaine semi-supervisée. Dans Laurent

Bougrain, éditeur, *Conférence Francophone sur l'Apprentissage Automatique*, pages 111–126, 2012. (Cité page 126.)

[Morvant *et al.*, 2012b] E. Morvant, A. Habrard, et S. Ayache. Parsimonious Unsupervised and Semi-Supervised Domain Adaptation with Good Similarity Functions. *Knowledge and Information Systems*, 33(2) :309–349, 2012. DOI : 10.1007/s10115-012-0516-7. (Cité page 126.)

[Morvant *et al.*, 2012c] E. Morvant, S. Koço, et L. Ralaivola. PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. Dans *Proceedings of International Conference on Machine Learning*, pages 815–822. Omnipress, 2012. (Full Paper, Acceptance rate : 27%). (Cité page 102.)

[Nock *et al.*, 2003] R. Nock, M. Sebban, et D. Bernard. A simple locally adaptive nearest neighbor rule with application to pollution forecasting. *Proceedings of International Journal of Pattern Recognition and Artificial Intelligence*, 17(8) :1369–1382, 2003. (Cité pages 27, 83 et 86.)

[Opelt *et al.*, 2004] A. Opelt, M. Fussenegger, A. Pinz, et P. Auer. Weak hypotheses and boosting for generic object detection and Recognition. Dans *Proceedings of European Conference on Computer Vision (ECCV)*, pages 71–84, 2004. (Cité page 90.)

[Pan et Yang, 2010] S. J. Pan et Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10) :1345–1359, 2010. (Cité page 36.)

[Parrado-Hernández *et al.*, 2012] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, et S. Sun. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13 :3507–3531, 2012. (Cité pages 58, 62, 65 et 100.)

[Paulsen, 2002] V. I. Paulsen. *Completely bounded maps and operator algebras*. Cambridge studies in advanced mathematics. Cambridge University Press, 2002. (Cité page 107.)

[Peterson et Jr., 1972] W. Wesley Peterson et E. J. Weldon Jr.. *Error-Correcting Codes*. The MIT Press, Cambrige, MA, 1972. (Cité page 121.)

[Quinlan, 1993] J. R. Quinlan. *C4. 5 : programs for machine learning*, volume 1. Morgan kaufmann, 1993. (Cité page 2.)

[Ralaivola, 2012] L. Ralaivola. Confusion-based online learning and a passive-aggressive scheme. Dans *Proceedings of Annual Conference on Neural Information Processing Systems*, pages 3293–3301, 2012. (Cité page 102.)

[Ralaivola *et al.*, 2010] L. Ralaivola, M. Szafranski, et G. Stempfel. Chromatic PAC-Bayes bounds for non-iid data : Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11 :1927–1956, 2010. (Cité page 62.)

[Ripley, 2007] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007. (Cité page 26.)

[Rosenblatt, 1958] F. Rosenblatt. The perceptron. *Psych. Rev*, 65(6) :386–408, 1958. (Cité page 2.)

[Roy *et al.*, 2012] S. D. Roy, T. Mei, W. Zeng, et S. Li. Socialtransfer : cross-domain transfer learning from social streams for media applications. Dans *Proceedings of ACM Multimedia Conference*, pages 649–658, 2012. (Cité page 36.)

[Saenko *et al.*, 2010] K. Saenko, B. Kulis, M. Fritz, et T. Darrell. Adapting visual category models to new domains. Dans *ECCV 2010*, pages 213–226. Springer, 2010. (Cité page 36.)

[Schölkopf et Smola, 2002] B. Schölkopf et A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization and beyond*. the MIT Press, 2002. (Cité page 27.)

[Seah *et al.*, 2010] C.W. Seah, I. Tsang, Y.S. Ong, et K.K. Lee. Predictive distribution matching svm for multi-domain learning. Dans *Proceedings of European Conference on Machine Learning and Principles of Data Mining and Knowledge Discovery*, volume 6321 de *LNCS*, pages 231–247. Springer, 2010. (Cité page 147.)

[Seeger, 2002] M. Seeger. PAC-bayesian generalization error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3 :233–269, 2002. (Cité pages 61, 63, 106 et 110.)

[Seldin *et al.*, 2012] Y. Seldin, N. Cesa-Bianchi, P. Auer, F. Laviolette, et J. Shawe-Taylor. PAC-Bayes-bernstein inequality for martingales and its application to multiarmed bandits. *Journal of Machine Learning Research - Proceedings Track*, 26 :98–111, 2012. (Cité page 63.)

[Seldin *et al.*, 2011] Y. Seldin, N. Cesa-Bianchi, F. Laviolette, P. Auer, J. Shawe-Taylor, et J. Peters. Pac-bayesian analysis of the exploration-exploitation trade-off. *arXiv preprint arXiv :1105.4585*, 2011. (Cité pages 176 et 178.)

[Seldin et Tishby, 2010] Y. Seldin et N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11 :3595–3646, 2010. (Cité page 63.)

[Senkene et Tempel'man, 1973] É. Senkene et A. Tempel'man. Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 13(4) :665–670, 1973. (Cité pages 29 et 177.)

[Settles, 2010] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010. (Cité page 176.)

[Shimodaira, 2000] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2) :227–244, 2000. (Cité pages 39 et 41.)

[Smeaton *et al.*, 2009] A. Smeaton, P. Over, et W. Kraaij. High-level feature detection from video in TRECVid : a 5-year retrospective of achievements. Dans *Multimedia*

Content Analysis, Theory and Applications, pages 151–174. Springer Verlag, 2009. (Cité page 146.)

[Smeaton *et al.*, 2006] A. F. Smeaton, P. Over, et W. Kraaij. Evaluation campaigns and TRECVid. Dans *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press. (Cité page 213.)

[Snoek *et al.*, 2005] C. Snoek, M. Worring, et A. W. M. Smeulders. Early versus late fusion in semantic video analysis. Dans *Proceedings of ACM Multimedia Conference*, pages 399–402, 2005. (Cité page 92.)

[Sutton et Barto, 1998] R. S. Sutton et A. G. Barto. *Reinforcement learning : An introduction*. Cambridge Univ Press, 1998. (Cité page 178.)

[Torralba et Efros, 2011] A. Torralba et A. A. Efros. Unbiased look at dataset bias. Dans *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011. (Cité page 4.)

[Tropp, 2011] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, pages 1–46, 2011. (Cité pages 101, 104, 106 et 120.)

[Urner *et al.*, 2011] R. Urner, S. Shalev-Shwartz, et S. Ben-David. Access to unlabeled data can speed up prediction time. Dans *Proceedings of International Conference on Machine Learning*, pages 641–648, 2011. (Cité page 43.)

[Valiant, 1984] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27 :1134–1142, 1984. (Cité page 18.)

[Vapnik, 1982] V. Vapnik. *Estimation of Dependences Based on Empirical Data : Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., 1982. (Cité page 19.)

[Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*. Springer, 1998. (Cité page 140.)

[Vapnik, 1995] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. (Cité page 18.)

[Vapnik et Chervonenkis, 1971] V. N. Vapnik et A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2) :264–280, 1971. (Cité page 19.)

[Wang et Kankanhalli, 2010] X. Wang et M. S. Kankanhalli. Portfolio theory of multimedia fusion. Dans *Proceedings of ACM Multimedia Conference*, pages 723–726, 2010. (Cité page 93.)

[Weinberger et Saul, 2009] K. Q. Weinberger et L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10 :207–244, juin 2009. (Cité pages 27 et 87.)

[Wickramaratna *et al.*, 2001] J. Wickramaratna, S. Holden, et B. Buxton. Performance degradation in boosting. Dans *Multiple Classifier Systems*, volume 2096 de *Lecture Notes in Computer Science*, pages 11–21. Springer Berlin Heidelberg, 2001. (Cité page 92.)

[Wolpert, 1992] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2) :241–259, 1992. (Cité page 92.)

[Xu *et al.*, 2012] H. Xu, C. Caramanis, et S. Mannor. Sparse algorithms are not stable : A no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1) :187–193, 2012. (Cité page 22.)

[Xu et Mannor, 2010] H. Xu et S. Mannor. Robustness and generalization. Dans *Proceedings of Annual Conference on Computational Theory*, pages 503–515, 2010. (Cité pages 22, 23, 24, 43, 128, 130, 138 et 202.)

[Xu et Mannor, 2012] H. Xu et S. Mannor. Robustness and generalization. *Machine Learning*, 86(3) :391–423, 2012. (Cité pages 22, 23, 24, 43, 128 et 130.)

[Yang et Jin, 2006] L. Yang et R. Jin. Distance Metric Learning : A Comprehensive Survey. Rapport Technique, Department of Computer Science and Engineering, Michigan State University, 2006. (Cité page 27.)

[Yao et Doretto, 2010] Y. Yao et G. Doretto. Boosting for transfer learning with multiple sources. Dans *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 1855–1862, 2010. (Cité page 40.)

[Yue *et al.*, 2007] Y. Yue, T. Finley, F. Radlinski, et T. Joachims. A support vector method for optimizing average precision. Dans *Proceedings of Special Interest Group on Information Retrieval conference*, pages 271–278, 2007. (Cité page 95.)

[Zadrozny, 2004] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. Dans *Proceedings of the international conference on Machine learning*, page 114. ACM, 2004. (Cité page 39.)

[Zhang, 2004] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5 :1225–1251, 2004. (Cité page 94.)

[Zhang et Yeung, 2010] Y. Zhang et D.-Y. Yeung. Transfer metric learning by learning task relationships. Dans *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1199–1208. ACM, 2010. (Cité page 176.)

[Zhong *et al.*, 2010] E. Zhong, W. Fan, Q. Yang, O. Verscheure, et J. Ren. Cross validation framework to choose amongst models and datasets for transfer learning. Dans *Proceedings of European Conference on Machine Learning and Principles of Data Mining and Knowledge Discovery*, volume 6323 de *LNCS*, pages 547–562. Springer, 2010. (Cité pages 55 et 132.)

[Žliobaitė, 2010] I. Žliobaitė. Learning under concept drift : an overview. *arXiv preprint arXiv :1010.4784*, 2010. (Cité page 175.)

Ce document a été préparé à l'aide de l'éditeur de texte GNU Emacs et du logiciel de composition typographique LaTeX 2_ε. Le style du document est librement inspiré du modèle de thèse élaboré par Julien Chiquet.

Titre Apprentissage de vote de majorité pour la classification supervisée et l'adaptation de domaine : approches PAC-Bayésiennes et combinaison de similarités

Résumé De nos jours, avec l'expansion d'Internet, l'abondance et la diversité des données accessibles qui en résulte, de nombreuses applications requièrent l'utilisation de méthodes d'apprentissage automatique supervisé capables de prendre en considération différentes sources d'informations. Par exemple, pour des applications relevant de l'indexation sémantique de documents multimédia, il s'agit de pouvoir efficacement tirer bénéfice d'informations liées à la couleur, au texte, à la texture ou au son des documents à traiter. La plupart des méthodes existantes proposent de combiner ces informations multimodales, soit en fusionnant directement les descriptions, soit en combinant des similarités ou des classifieurs, avec pour objectif de construire un modèle de classification automatique plus fiable pour la tâche visée. Ces aspects multimodaux induisent généralement deux types de difficultés. D'une part, il faut être capable d'utiliser au mieux toute l'information *a priori* disponible sur les objets à combiner. D'autre part, les données sur lesquelles le modèle doit être appliqué ne suivent nécessairement pas la même distribution de probabilité que les données utilisées lors de la phase d'apprentissage. Dans ce contexte, il faut être à même d'adapter le modèle à de nouvelles données, ce qui relève de l'adaptation de domaine. Dans cette thèse, nous proposons plusieurs contributions fondées théoriquement et répondant à ces problématiques. Une première série de contributions s'intéresse à l'apprentissage de votes de majorité pondérés sur un ensemble de votants dans le cadre de la classification supervisée. Ces contributions s'inscrivent dans le contexte de la théorie PAC-Bayésienne permettant d'étudier les capacités en généralisation de tels votes de majorité en supposant un *a priori* sur la pertinence des votants. Notre première contribution vise à étendre un algorithme récent, MinCq, minimisant une borne sur l'erreur du vote de majorité en classification binaire. Cette extension permet de prendre en compte une connaissance *a priori* sur les performances des votants à combiner sous la forme d'une distribution alignée. Nous illustrons son intérêt dans une optique de combinaison de classifieurs de type plus proches voisins, puis dans une perspective de fusion de classifieurs pour l'indexation sémantique de documents multimédia. Nous proposons ensuite une contribution théorique pour des problèmes de classification multiclasse. Cette approche repose sur une analyse PAC-Bayésienne originale en considérant la norme opérateur de la matrice de confusion comme mesure de risque. Notre seconde série de contributions concerne la problématique de l'adaptation de domaine. Dans cette situation, nous présentons notre troisième apport visant à combiner des similarités permettant d'inférer un espace de représentation de manière à rapprocher les distributions des données d'apprentissage et des données à traiter. Cette contribution se base sur la théorie des fonctions de similarités (ϵ,γ,τ)-bonnes et se justifie par la minimisation d'une borne classique en adaptation de domaine. Pour notre quatrième et dernière contribution, nous proposons la première analyse PAC-Bayésienne appropriée à l'adaptation de domaine. Cette analyse se base sur une mesure consistante de divergence entre distributions permettant de dériver une borne en généralisation pour l'apprentissage de votes de majorité en classification binaire. Elle nous permet également de proposer un algorithme adapté aux classifieurs linéaires capable de minimiser cette borne de manière directe.

Mots-clés Apprentissage Automatique, Vote de majorité, Théorie PAC-Bayésienne, Classification supervisée, Adaptation de domaine.

Title Learning Majority Vote for Supervised Classification and Domain Adaptation : PAC-Bayesian Approaches and Similarity Combination

Abstract Nowadays, due to the expansion of the web a plenty of data are available and many applications need to make use of supervised machine learning methods able to take into account different information sources. For instance, for multimedia semantic indexing applications, one have to efficiently take advantage of information about color, textual, texture or sound sources of the document. Most of the existing methods try to combine these multimodal informations, either by directly fusionning the descriptors or by combining similarities or classifiers, in order to produce a classification model more reliable for the considered task. Usually, these multimodal facets imply two main issues. On the one hand, one have to be able to correctly make use of all the *a priori* information available. On the other hand, the data, on which the model will be applied, does not come from the same probability distribution than the data used during the learning step. In this context, we have to adapt the model on new data, which is known as domain adaptation. In this thesis, we propose several theoretically-founded contributions for tackle these issues. A first serie of contributions studies the problem of learning a weighted majority vote over a set of voters in a supervised classification setting. These results fall within the context of the PAC-Bayesian theory allowing to derive generalization abilities for such a vote by assuming an *a priori* on the relevance of the voters. Our first contribution aims at extending a recent algorithm, MinCq, minimizing a bound over the error of the majority vote in binary classification. This extension can take into account an *a priori* belief on the performances of the voters. This belief is expressed as an aligned distribution. We illustrate its usefulness for combining nearest neighbor classifiers, and for classifier fusion on a multimedia semantic indexing task. Then, we propose a theoretical contribution for multiclass classification tasks. Our approach is based on an original PAC-Bayesian analysis considering the operator norm of the confusion matrix as an error measure. Our second series of contributions relates to domain adaptation. In this situation we present our third result for combining similarities in order to infer a representation space for moving closer the learning distribution and the testing distribution. This contribution is based on the theory of learning from (ϵ,γ,τ)-good similarity functions and is justified by the minimization of an usual bound in domain adaptation. For our last contribution, we propose the first PAC-Bayesian analysis for domain adaptation. This analysis is based on a consistent divergence measure between distributions allowing us to derive a generalization bound for learning majority votes in binary classification. Moreover, we propose a first algorithm specialized to linear classifiers and able to directly minimize our bound.

Keywords Machine Learning, Majority vote, PAC-Bayesian theory, Supervised classification, Domain Adaptation.